Université de Sherbrooke

L'utilisation d'annotation génomique est un paramètre expérimental en bio-informatique

Par
Joël Simoneau
Programme de Biochimie

Mémoire présenté à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de maître ès sciences (M. Sc.)
en Biochimie

Sherbrooke, Québec, Canada
Février, 2020

Membres du jury d'évaluation

Michelle S. Scott Ph.D.,
Faculté de Médecine et des Sciences de la Santé,
Département de Biochimie et de Génomique Fonctionnelle

Ryan Gosselin ing., Ph.D.,
Faculté de Génie,
Département de Génie Chimique et de Génie Biotechnologique

Guylain Boissonneault Ph.D.,
Faculté de Médecine et des Sciences de la Santé,
Département de Biochimie et de Génomique Fonctionnelle

Pierre-Étienne Jacques Ph.D.,
Faculté des Sciences,
Département de Biologie

Sommaire

L'utilisation d'annotation génomique est un paramètre expérimental en bio-informatique

Par
Joël Simoneau
Programme de Biochimie

Mémoire présenté à la Faculté de médecine et des sciences de la santé en vue de l'obtention du diplôme de maître ès sciences (M. Sc.) en Biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

La biologie moléculaire est l'étude des mécanismes du vivant, à l'échelle des molécules. Les capacités d'une espèce vivante sont définies dans son matériel génétique. Ainsi, l'étude du génome d'une espèce devrait être la clef à la compréhension de son fonctionnement. Les diverses molécules fonctionnelles sont produites à partir d'information contenue dans le génome, sous forme d'ARN et de protéines. À la suite de l'obtention de la première séquence complète d'un génome humain, il s'est révélé la nécessité d'en étudier le contenu, et donc d'en identifier les divers produits encodés.

Les annotations génomiques sont des références et des ressources qui contiennent l'information des divers produits d'un génome. Il s'agit à la fois d'un lieu d'accumulation d'information, et d'un outil de référence dans des analyses bio-informatiques. Plusieurs projets d'annotation existent parallèlement, chacun s'intéressant à annoter le même génome de référence.

Le séquençage de l'ARN est une technologie qui a révolutionné l'étude de la biologie moléculaire en permettant l'identification et la quantification de molécules d'ARN de manière hautement parallèle et à faible coût. La transformation des données machines issue d'une expérience de séquençage en données biologiquement interprétables requiert l'utilisation d'un pipeline bio-informatique. En plus de divers outils logiciels nécessaires, une annotation génomique doit être utilisée pour définir les gènes et leurs produits étudiés.

Ce mémoire argumente que les annotations génomiques ne sont pas historiquement considérées comme étant méthodologiquement importantes, tout en démontrant que le choix d'annotation génomique introduit un biais de quantification important en séquençage de l'ARN.

Mots-clés: RNA-seq, Annotation génomique, Reproductibilité, Méthodes

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ADN Acide déoxyribonucléique

ADNc ADN complémentaire

APPRIS *Annotation of principal and alternative splice isoforms*

ARN Acide ribonucléique

ARNm ARN messager

ARNr ARN ribosomal

ARNt ARN de transfert

EBI *European Bioinformatics Institute*

CCDS *Concensus CDS*

CDS Séquence codante, *Coding sequence*

ENCODE *Encyclopedia Of DNA Elements*

EST Marqueur de séquence exprimée, *Expressed Sequence Tag*

GFF *General Feature Format*

GTF *Gene Transfer Format*

HGNC *HUGO Gene Nomenclature Committee*

HGP Projet Génome Humain, *Human Genome Project*

HUGO *Human Genome Organisation*

Iso-seq Séquençage d'isoforme, *Isoform Sequencing*

NIH *National Institutes of Health*

NGS Séquençage de nouvelle génération, *Next-Generation Sequencing*

nt nucléotide

PTM Modifications post-transcriptionnelle, *Post-Transcriptional Modifications*

RNA-seq Séquençage de l'ARN, *RNA sequencing*

snoRNA petit ARN nucléolaire, *small nucleolar RNA*

snRNA petit ARN nucléaire, *small nuclear RNA*

TSL *Transcript Support Level*

UCSC *University of California Santa Cruz*

UTR Région non traduite, *Untranslated region*

VEGA *Vertebrate and Genome Annotation*

Remerciements

Le passage du baccalauréat aux études supérieures n'est pas une transition sans heurt. Il faut apprendre à réfléchir la science différemment, en tant que personne actrice et non observatrice. La science c'est prendre action vers l'inconnu, ce qui amène son lot de problématiques ouvertes, de chemins ardemment parcourus mais parfois abandonnées, de remises en question scientifiques et personnelles. Mais la science c'est aussi une aventure parcourue à plusieurs, des discussions animées, des réalisations partagées et des succès à célébrer.

Je me dois de remercier Michelle Scott et Ryan Gosselin qui ont accepté de m'accompagner dans cette aventure scientifique et humaine. Vous m'avez donné la liberté d'explorer, tout en étant disponibles pour m'épauler à travers les difficultés et les incompréhensions. Vous m'avez apporté des questionnements complémentaires et pertinents qui ont solidifié ma démarche et ma recherche. Je me suis senti supporté et valorisé, et je suis reconnaissant des opportunités que vous m'avez offertes. J'avais tendance à universaliser mon expérience de maîtrise, mais je réalise de plus en plus que les réalités sont diverses et complexes, et que je suis choyé d'avoir eu l'expérience qui est la mienne.

J'aimerais aussi remercier Guylain Boissonneault et Pierre-Étienne Jacques d'avoir accepté de siéger sur mon comité d'évaluation, mais aussi des interactions positives et intéressées que l'on a partagées le long de mon parcours. J'espère que vous en apprécierez la lecture !

Je remercie aussi les personnes que j'ai côtoyées professionnellement et amicalement tout le long de ces deux dernières années. Je pense aux moments de discussions et d'entraide au laboratoire, aux gâteaux d'anniversaire, aux bières du vendredi soir, aux différents moments d'échange scientifique. Je pense à la camaraderie et au soutien que l'on se partage. Ce mémoire est une autre étape dans mon parcours, et je dois maintenant dire au revoir à bien des gens. Je vous dis à la prochaine fois.

Chapitre 1
Introduction

Contrairement à ce que l'on pourrait penser, la bio-informatique n'a pas été originellement définie comme l'étude de la biologie par méthodes informatique, bien qu'elle est dorénavant conceptualisée comme telle. Le terme bio-informatique provient de «l'étude de processus informatique des systèmes biotiques» (Hogeweg, 2011), où processus informatique est défini comme l'étude de la capacité à stocker, communiquer et utiliser de l'information. Il est aussi élégant de constater que la frontière du réalisable en bio-informatique est directement reliée au développement du matériel informatique utilisé (Spengler, 2000). Ainsi, notre capacité à comprendre l'informatique des systèmes biotiques est limitée par nos propres systèmes informatiques.

L'informatique biotique s'articule autour de trois polymères biologiques, l'ADN, l'ARN et les protéines. Selon la théorie éculée du dogme central de la biologie moléculaire, ces trois types de molécules sont respectivement responsables des étapes principales de stockage, communication et utilisation de l'information biologique pertinente aux fonctions du vivant (Crick, 1970). Il a été mis en lumière depuis lors que les interrelations et les rôles des différentes molécules sont beaucoup plus complexes qu'initialement énoncées, mais il n'en reste que leurs rôles centraux sont conservés, et diversifiés (Camacho, 2019).

Une des promesses majeures de la bio-informatique à l'étude de la biologie moléculaire est la capacité à interroger très largement, et préférentiellement sans biais, des sous-ensembles de molécules. Alors que la biologie de laboratoire utilise plutôt un schéma de recherche gène centrique, issu de contraintes expérimentales, la bio-informatique, appuyée par une instrumentation et des protocoles expérimentaux spécialisés, promet théoriquement l'analyse parallèle de tous les gènes et leurs produits. Cette promesse tient sur l'hypothèse que nous connaissons la diversité biologique et que le développement de nos outils bio-informatiques et de nos références biologiques prend en considération cette diversité.

Les techniques bio-informatiques d'analyse ont souvent recours à des bases de données contenant des informations diverses sur les gènes ou molécules étudiées (e.g. séquence, structure, propriété physicochimique) qui sont intégrées par les modèles utilisés. Ces bases de données sont très diverses en termes d'information contenue, de méthode de récolte

d'information, de fréquence de mise à jour, d'entité responsable, etc. (Cannata, Merelli, & Altmare, 2005). L'importance des données de référence dans une analyse est aussi dépendante de la manière dont les données sont utilisées dans le modèle, et cela peut avoir une vaste gamme d'impact.

Ce mémoire s'interroge sur l'utilisation des annotations génomiques, représentant des bases de données contenant la position et la structure des gènes dans le génome, dans le contexte de la quantification d'ARN dans la technologique du séquençage de l'ARN (RNA-seq). Le RNA-seq est un choix qui s'impose, puisque c'est la technologie à haut débit centrale à l'étude du transcriptome, et la technologie qui a été adaptée en une multitude de techniques différentes qui élargissent considérablement la boîte à outils de la biologie moléculaire (Hrdlickova, Toloue & Tian, 2017). Le chapitre premier présente le contexte technologique et scientifique concernant la création des annotations génomiques, l'information contenue dans celle-ci et leur utilisation en RNA-seq. Le chapitre second présente une analyse publiée dans *Briefings in Bioinformatics* de la méthodologie rapportée par les personnes utilisatrices de RNA-seq dans la littérature. Le chapitre troisième présente une analyse en révision chez *NAR Genomics and Bioinformatics* de l'impact des annotations génomiques, en relation avec d'autres choix méthodologiques, dans le RNA-seq. Le dernier chapitre revient sur les éléments clés, illustre les problématiques potentielles et identifie les prochaines étapes à suivre.

1.1 Les références biologiques

Les bases de données de références biologiques agissent en même temps comme lieux d'accumulation de connaissances et outils de recherche scientifique. Ces bases de données sont diverses en nature, mais très souvent organisées par espèce. Les données disponibles, et leur qualité, varient grandement en fonction de l'espèce étudiée (Lewin et al., 2018), ce qui signifie que l'étude des références biologiques doit se définir par espèce. Dans les travaux présentés dans ce mémoire, nous avons utilisé *Homo sapiens* considérant l'anthropocentrisme de la recherche scientifique. Avec des projets comme le séquençage du génome de 2636 personnes islandaises (Gudbjartsson et al., 2015), et avec l'annonce du UK BioBank de séquencer 500 000 individus (Sudlow et al., 2015), l'*Homo sapiens* est l'espèce possédant le plus de données d'individus séquencés.

1.1.1 Le génome

Le Projet Génome Humain (HGP, Human Genome Project), soit le séquençage du premier génome humain, est considéré comme étant le mégaprojet de la biologie moléculaire. Bien que son utilité apparente, en relation avec le coût exorbitant de 3 milliards de dollars américains du projet, ait été mise en doute initialement, le HGP a complètement transformé la recherche en biologie de laboratoire (Hood & Rowen, 2013). En plus d'élucider, en très grande partie, la séquence d'un génome humain type, le HGP a été un projet phare en termes de collaboration internationale pour la biologie, mais aussi d'utilisation et de développement de logiciel libre, de publication de données ouvertes (International Human Genome Sequencing Consortium, 2001) et de développement de technologie de séquençage à haut débit (Hood & Galas, 2003).

Aujourd'hui, le génome humain utilisé dans les analyses bio-informatiques se dit génome de référence. L'information contenue dans le génome de référence se lit comme une carte, où chaque position est référée par un système de coordonnées qui comprend le nom du chromosome, la position le long du chromosome et le brin et où chaque position contient l'information d'un seul nucléotide d'ADN, soit A, T, G ou C. Ce génome de référence occupe une place essentielle dans la plupart des analyses bio-informatiques considérant qu'il s'agit de la source première de n'importe quelle séquence d'ADN et d'ARN.

Bien que la première ébauche du génome ait été annoncée en 2001 (International Human Genome Sequencing Consortium, 2001), puis sa version complété en 2004 (International Human Genome Sequencing Consortium, 2004), le génome de référence est toujours incomplet, et évolue dans le temps (Guo et al., 2017). Il convient de surpréciser le terme génome de référence, car il ne correspond pas exactement à ce que l'on pourrait considérer comme étant la séquence génétique d'un individu. Le génome de référence humain est une construction haploïde, avec des loci alternatifs, alors qu'un individu possède un génome diploïde (Church et al., 2011). Maintenant qu'il existe de plus en plus de données de séquençage de génomes complets, il y a des appels à revoir notre méthode de construction du génome de référence, en tentant d'utiliser à tout coup les allèles majoritaires présents dans la population, ou bien de définir plusieurs génomes de référence qui représenteraient des archétypes divergents, mais largement retrouvés dans la population (Ballouz, Dobin, & Gillis, 2019). Qui plus est, on imaginerait le génome de référence comme étant un échantillon type sain. Il est cependant très difficile d'établir ce qu'est un individu sain, et des analyses du génome de référence le décrive comme étant un individu à haut risque de diabète de type 1 et d'hypertension (Chen & Butte, 2011).

L'obtention d'un génome complet se voulait un élément clé de réponse à des questions fondamentales sur la biologie humaine, soit le nombre de gènes, leurs rôles, et dans un objectif plus médical, l'identification de mutations et leur lien avec des maladies (Collins, Green, Guttmacher, & Guyer, 2003). Le génome humain se devait d'être la pierre d'assise de la médecine moderne basée sur la génomique, mais il s'est avéré peu interprétable en lui-même. Une annotation génomique en bonne et due forme était donc nécessaire pour donner au génome toute sa valeur en tant qu'outil de découverte (Stein, 2001).

1.1.2 Les annotations génomiques

Le génome de référence humain n'est qu'un ensemble de plus de trois milliards de nucléotides ordonnés sur 24 différents chromosomes, le chromosome mitochondrial et quelques locis alternatifs (International Human Genome Sequencing Consortium, 2004). Les annotations génomiques sont les outils d'interprétation de ces séquences, outils qui doivent relier les séquences à un sens biologique (Stein, 2001). Le terme annotation est largement utilisé en biologie et décrit simplement un outil d'information. Il existe plusieurs niveaux d'annotation qui s'intéressent à différents types d'information (i.e. structurelle, fonctionnelle, interaction). Dans cet ouvrage, nous n'étudierons que les annotations nucléotidiques, soit celles qui s'intéressent à la définition des gènes, nommées annotations génomiques.

Pour identifier les gènes, il faut tout d'abord s'entendre sur la définition du terme. Cette tâche n'est pas triviale puisque la définition du terme gène est directement reliée à notre compréhension de son rôle biologique, et ce rôle est en constante précision et redéfinition (Hopkin, 2009). Originellement un terme abstrait qui individualisait la capacité d'héritabilité du vivant, le gène peut aujourd'hui être décrit comme une section du génome pouvant être transcrite en ARN fonctionnel. Volontairement imprécise, cette définition ne s'aventure pas à un degré de précision nécessaire pour en faire la base d'une structure de donnée.

Puisque les annotations génomiques sont aussi des ressources informatiques, l'information qui y est contenue doit être organisée selon une structure fixe qui sera connue et exploitée par d'autres outils et logiciels. Présentement, les annotations génomiques sont distribuées dans des fichiers GTF (*Gene Transfer Format*) ou GFF (*General Feature Format*), qui contiennent la même information, organisée différemment, sur la structure des gènes (Stein, 2013). L'information sur un gène est organisée hiérarchiquement. Un gène possède un ou

plusieurs transcrits, qui chacun possède un ou plusieurs exons. Les différents transcrits d'un même gène sont également nommés isoformes. Un transcrit peut aussi posséder une séquence nommée CDS (*Coding Sequence*), qui définit la région du transcrit pouvant être traduite en protéine. Chacun de ces éléments est défini en fonction d'une position de début et de fin sur un brin d'un chromosome, permettant ainsi d'en trouver la séquence en se référant au génome de référence (Brent, 2007).

Ainsi, la structure de l'information dans le format de fichier des annotations génomiques représente une série de contraintes et donc de définition du gène. Par exemple, puisque chaque transcrit ne peut avoir qu'au maximum une seule CDS, les ARN polycistroniques avec des séquences codantes chevauchantes ne peuvent y être représentés adéquatement (Brunet et al., 2019), bien que dorénavant étudiés chez *Homo sapiens* (Mouilleron, Delcourt, & Roucou, 2016). Un autre exemple serait la présence de modifications post-transcriptionnelle (*Post-Transcriptional Modifications*, PTM) sur les ARNs matures. Certains ARNs nécessitent des modifications chimiques de nucléotides spécifiques afin de pouvoir être fonctionnels. Ces modifications, par exemple la méthylation d'un résidu adénosine (A), peuvent être fréquentes et cruciales à la biologie des ARNs, mais cette information ne se retrouve pas dans les annotations génomiques courantes. Il est à noter que le RNA-seq actuel ne peut pas détecter de façon fiable les diverses modifications de nucléotides (Nachtergaele & He, 2017).

1.1.3 L'ARN

L'ensemble des ARN, des transcrits, d'une espèce est nommé le transcriptome. Il est possible d'obtenir un transcriptome à partir d'un génome en faisant l'extraction de toutes les séquences contenues dans une annotation génomique, avec l'hypothèse que cette dernière est complète. Alors que le génome est composé de plusieurs chromosomes ayant des caractéristiques similaires, le transcriptome est composé d'ARN qui possèdent une vaste gamme de caractéristiques et rôles biologiques différents. Les ARN sont usuellement classifiés selon un biotype, soit une classe indiquant leur fonction ou caractéristique principale identifiée. La section suivante présente un très bref aperçu non exhaustif de quelques types d'ARN qui sont mentionnés dans ce document.

Le premier niveau de classification usuellement utilisé est la répartition entre les ARN codants et non-codants. La propriété d'être codant est celle de posséder une CDS, soit une

séquence définissant une protéine. Les ARN codants sont aussi nommés ARNm (ARN messager), puisqu'ils portent un message d'importance qui est celui de la séquence des protéines, qui sont historiquement considérées comme la partie centrale fonctionnelle de la cellule (Brenner et al., 1961). Ainsi, la quantification des ARNm est parfois utilisée comme une variable substitutive pour l'étude la quantification protéique, bien qu'il ne soit pas possible d'établir un réel rapport quantitatif uniforme entre les deux (Gry et al., 2009). L'avantage de quantifier les ARN est une beaucoup plus grande uniformité chimique des différentes molécules, contrairement aux protéines qui sont beaucoup plus diverses, et donc difficile à isoler et étudier sans biais.

Outre les ARNm, une vaste gamme croissante d'autres biotypes d'ARN existe, ensemble aussi connu sous le nom d'ARN non-codant. Historiquement, les ARN ribosomaux (ARNr) et les ARN de transfert (ARNt) sont les deux premiers biotypes d'ARN non-codant étudiés et caractérisés, où leur rôle en relation avec la production de protéines à partir d'ARNm est mis en évidence (Dennis, 1972). Les ARNr sont les principaux constituant des ribosomes, usine de traduction des ARNm en protéines, et représentent aussi la très grande majorité des ARN cellulaires en termes de quantité de séquence (Palazzo et al., 2015). Les ARNt ont le rôle de reconnaissance des codons présents dans les CDS des ARNm, et participent donc à la synthèse protéique en amenant les bons acides aminés aux ribosomes (Mitra, 1978). Les ARNt sont les ARN les plus abondants en termes de nombre de molécules (Palazzo et al., 2015). D'autres petits ARN hautement structurés, comme les petits ARN nucléaires (snRNA) et nucléolaires (snoRNA) occupent aussi des rôles importants dans la régulation et maturation des ARN (Dupuis-Sandoval et al., 2015). Alors que les biotypes d'ARN non-codant jusqu'à présent énoncés sont caractérisés par des rôles et structures conservés et définis, d'autres biotypes regroupent des classes beaucoup plus hétérogènes et cryptiques. Les longs ARN non-codants et les pseudogènes sont deux biotypes qui regroupent des ARN caractérisés par l'absence de CDS, une certaine longueur minimum et l'absence de structure conservée. Les pseudogènes se démarquent des longs ARN non-codants par une ressemblance individuelle à des gènes codants, avec l'hypothèse que les pseudogènes sont des copies dégénérées, principalement sans fonction connue, de gènes codants (Pei et al., 2012).

1.2 Séquençage de l'ARN (RNA-seq)

Le séquençage de l'ARN (RNA-seq) est une technologie permettant l'identification et la quantification de molécules d'ARN extraites d'un échantillon biologique. Le terme RNA-seq est usuellement utilisé pour décrire la seconde génération de technologies de séquençage, le *Next-Generation Sequencing* (NGS), malgré que le terme plus général décrit seulement la capacité d'identifier des molécules d'ARN. Afin d'insérer le RNA-seq dans son contexte technologique, la section suivante présente un bref survol de l'évolution des générations de technologies de séquençage, ainsi que les détails entourant un pipeline type bio-informatique nécessaire au RNA-seq.

1.2.1 Les technologies de séquençage

Les technologies de séquençage sont présentement classifiées selon trois générations différentes, où la troisième est toujours en définition et développement. La première génération est simplement caractérisée par la nouvelle capacité de séquencer des brins d'ADN (Sanger & Coulson, 1975). La seconde génération est supportée par une série d'innovations technologiques permettant une drastique augmentation de la quantité d'information séquençable en un laps de temps beaucoup plus rapide, réduisant ainsi significativement le coût de séquençage, et augmentant son utilisation en recherche (Heather & Chain, 2016). La troisième génération, toujours en développement, s'oriente principalement vers la capacité à séquencer des molécules individuelles, enlevant la nécessité d'amplification du matériel génétique des générations précédentes (Heather & Chain, 2016).

Ces technologies, sauf quelques exceptions, sont en fait des technologies de séquençage de l'ADN (Ozsolak et al., 2009). Dans le cas où nos séquences à identifier sont constituées d'ARN, il est possible d'en effectuer la transcription inverse des brins d'ARN en ADN avec une transcriptase inverse, puis de séquencer les nouveaux brins.

Première génération. La première génération de technologie de séquençage est nommée le séquençage de type Sanger, développé par le scientifique éponyme (Sanger & Coulson, 1975). Cette technologie se base sur le concept de complémentarité de l'ADN, et de l'utilisation d'une ADN polymérase, pour identifier la séquence nucléotidique d'un brin d'ADN. L'ADN est un polymère constitué de quatre monomères (A, T, C et G), aussi appelés nucléotides. Ces monomères sont liés linéairement par liens covalents en brin d'ADN.

Chaque nucléotide peut aussi s'apparier à un autre nucléotide par des liaisons hydrogène, où chaque nucléotide possède un partenaire de liaison préférentiel, soit A avec T et C avec G. La structure de l'ADN, soit la double hélice, est formée de deux brins complémentaires d'ADN. Lors de la réplication de l'ADN, une enzyme nommée ADN polymérase utilise la séquence contenue d'un des deux brins, appelé brin matrice, pour recréer un brin complémentaire. Le concept de la technologie de séquençage par synthèse, tel le séquençage de type Sanger, est basé sur l'observation du processus de création d'un brin complémentaire, qui permet d'obtenir la séquence du brin matrice. La stratégie classique de séquençage pour un brin d'ADN implique l'utilisation de nucléotides modifiés, qui sont additionnés de fluorochrome ou radiomarqués afin de pouvoir les identifier individuellement. Ces nucléotides modifiés empêchent aussi l'ajout de nucléotides subséquents. En utilisant plusieurs clones du brin d'ADN à séquencer, des ADN polymérases et un mélange de nucléotides modifiés et naturels, il est possible de synthétiser des brins complémentaires de plusieurs longueurs différentes, qui vont tous finir par un nucléotide marqué (Sanger, Nicklen et Coulson, 1977). En séparant ces brins par taille et en identifiant la nature des nucléotides marqués, il est possible de reconstituer la séquence originale du brin d'ADN d'intérêt. Au fil des années, plusieurs implémentations technologiques ont optimisé le rendement de la méthode, par exemple la séparation par taille par électrophorèse sur capillaire avec détection automatique des bases, au lieu de migration sur gel avec détection manuelle (Halloran, Du et Wilson, 1993), mais il n'en reste que le séquençage Sanger est caractérisé par la création de multiples clones du brin matrice, avec toutes les diverses troncations de nucléotides possibles.

Seconde génération. Le séquençage de seconde génération a révolutionné la génomique et la transcriptomique par une modification au protocole de séquençage par synthèse de type Sanger. Au lieu de devoir générer toutes les troncations terminales possibles des brins d'ADN à identifier, la technologie de pyroséquençage, introduite en NGS, permet d'observer séquentiellement l'ajout de nucléotides sur des brins d'ADN. Lors de l'incorporation d'un nucléotide par l'ADN polymérase, un pyrophosphate est libéré. Cette molécule peut ensuite être catalysée en ATP par l'ATP sulfurylase, ATP qui sera consommée par une luciférase en présence de luciférine, ce qui émet une lumière visible, détectable par caméra (Nyrén, Pettersson et Uhlén, 1993). Le pyroséquençage est une approche pour éliminer des limitations techniques de la première génération, mais d'autres approches ont aussi été développées, tel que les nucléotides avec flurochrome clivable d'Illumina (van Dijk, 2014). Ainsi, le NGS se libère de plusieurs contraintes du séquençage de première génération, comme le clonage, la nécessité de nucléotides marqués radioactivement et une

large quantité nécessaire de réactif, ce qui augmente l'efficacité de séquençage et donc en réduit les coûts associés (Heather & Chain, 2016). Bien que le NGS a de plusieurs ordres de grandeur l'efficacité de nombre de nucléotides séquencés, le séquençage de type Sanger offre des lectures plus longues et avec une meilleure confiance (Margulies et al., 2005). Il n'en reste qu'en séquençant plusieurs fois les mêmes séquences, il est possible d'assembler les lectures en fragments plus longs avec les éléments se chevauchant, et de corriger les erreurs de séquences en même temps par méthode de consensus.

Troisième génération. La troisième génération est plus difficile à caractériser, puisqu'encore en développement. Les deux caractéristiques principales de la troisième génération sont des réponses aux limitations techniques des deux générations précédentes, soit éliminer la nécessité de la multiplication et fragmentation des données à séquencer (Heather & Chain, 2016). Par exemple, la technologie de Pacific Biosciences l'observation en temps réel d'un brin unique d'ADN qui se fait séquencer son brin complémentaire par une ADN polymérase (Eid et al., 2009). Théoriquement, l'atteinte technologique de la possibilité de séquencer directement un brin d'ARN, sans en faire la fragmentation, la multiplication ni la transcription inverse, et ce à haut débit, serait la panacée de tout projet d'annotation. Cela permettrait d'avoir une image réelle des ARN présents biologiquement, et donc retirait les problématiques d'inférence de molécules complètes à partir de fragments.

1.2.2 Pipeline de RNA-seq

Les premières expériences publiées de RNA-seq datent maintenant de plus d'une dizaine d'années (Mortazavi et al., 2008). Depuis ce temps, les caractéristiques des données recueillies ont évolué, et les méthodes expérimentales ont dû suivre cette progression. En observant les données de séquençage publiées librement, on observe une place de plus en plus grande du séquençage en paire, produisant des lectures plus longues, et un plus grand nombre de celles-ci (Van den Berge et al., 2019). Bien que les concepts centraux nécessaires au traitement de données de RNA-seq soient restés les mêmes au courant des années, l'évolution de nos connaissances, la découverte de biais expérimentaux et l'évolution des séquenceurs demandent une mise à jour quasi constante des approches utilisées. Nous détaillerons les principales étapes nécessaires à la génération et au traitement des données de RNA-seq. Le pipeline est séparé en deux grandes étapes, soit le *in vitro* et le *in silico*, décrivant respectivement les étapes avant et après le séquenceur.

Pipeline in vitro. Le pipeline *in vitro* consiste en l'extraction, la préparation et le séquençage de l'ARN (Sultan et al. 2014). Les différents choix méthodologiques à faire se doivent d'être en phase avec la question biologique qui est étudiée (Chao et al., 2019). L'extraction est l'isolation de l'ARN à séquencer provenant d'un échantillon cellulaire. Cela consiste aussi à subdiviser la population totale d'ARN car il est fort peu courant de faire un séquençage d'ARN total. En effet, les ARN ribosomaux constituent environ 90% des ARN cellulaires en termes de concentration, et cela peut obstruer la quantification des autres biotypes d'ARN. Les ARNr sont habituellement retirés en les dégradant spécifiquement, ou en isolant un certain biotype d'ARN. On peut conceptualiser le séquençage comme une technique d'échantillonnage aléatoire sans remise, où l'on s'attend à retrouver, pour un ARN d'intérêt, un pourcentage de lecture similaire à son pourcentage de nucléotides sur l'ensemble des nucléotides présents dans l'échantillon. Après les ARN sont fragmentés, afin de pouvoir obtenir des lectures uniformément distribuées sur le long des ARN puisque ceux-ci peuvent être beaucoup plus long que la longueur des lectures de séquençage (Wery et al., 2013). Puisque les technologies de seconde génération séquencent de l'ADN et non de l'ARN, il faut effectuer la transcription inverse de l'ARN en ADN dit complémentaire (ADNc). Suivant l'obtention d'ADNc, celui-ci est préparé pour être séquencé, bien souvent en utilisant une trousse provenant de la compagnie développant la technologie de séquençage utilisée. Cette préparation, dont le produit se nomme librairie, inclut souvent l'ajout de séquences d'index et d'amorces aux ADNc (Trombetta et al., 2015). Les libraires sont par la suite séquencées. En fonction des protocoles expérimentaux utilisés, le séquençage peut produire des lectures d'une variété de longueurs, principalement entre 50 et 200 nucléotides. Certains protocoles ne permettent pas de savoir de quel brin, sense ou antisense, les lectures proviennent. Une caractéristique clef du séquençage est aussi la possibilité de produire des lectures en paires, soit une lecture à partir de chaque extrémité des ARN, ce qui permet d'avoir une meilleure confiance en la séquence, ou de séquencer des fragments plus longs que la longueur de lecture. La préparation *in vitro* est caractérisée de reproductible, mais biaisée. L'utilisation des mêmes protocoles devrait fournir les mêmes résultats, mais il peut y avoir beaucoup de différences de résultats entre les divers protocoles (Nottingham et al., 2016).

Pipeline *in silico*. Le pipeline *in silico* consiste en la transformation des données machines en données interprétables, ici des quantifications relatives de gènes. Le point de départ est un fichier FASTQ, qui contient la séquence des lectures et un score de qualité. La première étape est le contrôle qualité des lectures, qui consiste en retirer les extrémités de lecture qui sont de mauvaise qualité, retirer les séquences non biologiques qui ont été

séquencées, soit des amorces ou index de la librairie, et retirer les lectures qui sont trop courtes (Williams et al., 2016). Suivant cela, il faut déterminer d'où proviennent les lectures. Pour ce faire, l'alignement tente de placer chaque lecture à la séquence biologique le plus susceptible de l'avoir généré. Ici, il est possible divers type d'alignement, soit génomique ou transcriptomique, donc soit en cherchant l'origine des lectures dans l'entièreté de la séquence du génome, ou parmi les transcrits décrits dans une annotation génomique. Une annotation génomique peut aussi être utilisée par certains logiciels d'alignement génomique afin de faciliter l'alignement de gènes possédant des séquences discontinues. Après l'alignement des lectures, celles-ci sont quantifiées. La quantification est l'acte d'identifier le prorata de lectures appartenant à divers objets d'annotation différents, usuellement soit des gènes ou des transcrits (Conesa et al., 2016).

1.3 Élaboration des annotations génomiques

Les différentes méthodes de construction des annotations génomiques peuvent être séparées sur un spectre, avec dans les extrêmes les méthodes de curation manuelles et les méthodes de prédiction computationnelle pures. Les différentes méthodes d'annotation ont accès à un ensemble de preuves similaires, mais la grande différence est le processus décisionnel qui est sur le continuum entre une personne, dans le cas de la curation manuelle, ou à une machine, dans le cas de la prédiction.

Lincoln Stein (2001) a classifié les types d'effort d'annotation comme de musée, de jamborée, d'industrie artisanale et d'usine. L'usine est la méthode la plus impersonnelle, mais la plus agile et reproductible, soit la prédiction par ordinateur. La majorité des données d'annotation provient des usines, soit des projets majeurs (comme le gene-build d'Ensembl décrit plus loin). Ce sont aussi les informations considérées comme les moins fidèles. L'usine est par contre une bonne méthode comme point de départ pour les trois autres types d'effort d'annotation. La méthode musée est l'annotation manuelle par des personnes curatrices professionnelles. Ces individus sont des biologistes qui parcourent la littérature et les données biologiques de séquençage pour compiler les informations connues sur les gènes. Les efforts de type musée englobent souvent plus que les annotations génomiques, touchant aussi à la fonction des gènes et des différents isoformes. L'industrie artisanale représente l'effort d'annotation porté à temps partiel par des groupes recherche qui soumettent ou corrigent des annotations concernant leurs gènes d'intérêt. Les jamborées sont aussi une forme d'annotation communautaire, où des biologistes et bio-informaticiens se regroupent

à travers des ateliers de travail, pour y construire et améliorer des annotations génomiques. Ces deux méthodes communautaires sont beaucoup plus anecdotiques, où ce sont souvent des organisations regroupées autour de modèles animaux d'intérêt qui poussent ces efforts, tel que par exemple le jamborée d'annotation de la drosophile, suivant le séquençage de son génome (Pennisi, 2000).

1.3.1 Données biologiques utiles

Les données biologiques utilisées pour élaborer les annotations produisent des annotations d'une vaste gamme de degrés de confiance. Les informations utilisées sont généralement des séquences de protéines, des EST (*Expressed Sequence Tags*) et des séquences d'ADNc (ADN complémentaire).

Séquences protéiques. À partir des séquences de protéines, il est possible d'inférer la séquence d'ARN messager correspondante et donc la séquence d'ADN initiale. Il est ensuite possible de chercher cette séquence d'ADN, avec ou sans écart causé par la présence d'intron, dans le génome afin d'en trouver la position. Puisque cette technique se base sur la séquence protéique, il est impossible d'obtenir de l'information sur le contexte nucléotidique environnant, et donc aucune information sur les séquences non traduites de l'ARN messager correspondant. Ainsi, on ne fait que trouver la séquence codante des gènes. Il est par contre possible de supplémenter cette méthode par des informations d'ADNc et de EST afin de pouvoir résoudre l'entièreté de la séquence des gènes.

Séquençage d'EST. Le séquençage de EST figure dans les premières techniques d'identification de gènes à haut débit. Pour séquencer des EST, il faut isoler des ARNm d'un extrait cellulaire, puis en séquencer les extrémités sans les fragmenter. Les EST sont souvent identifiés comme 5' EST si séquencés à partir du début des ARNm, ou 3' EST pour l'inverse. L'objectif de la production d'EST est l'obtention de séquences assez longues pour identifier les gènes ainsi que leurs frontières de début et de fin, tout en séquençant de manière conservatrice (Matsubara & Okubo, 1993). Les EST peuvent bien complémenter les séquences de protéines, puisque si les EST séquencés chevauchent les séquences de protéines trouvées précédemment, il est possible de reconstituer des ARNm complets. Les premières expériences de séquençage d'EST précèdent les projets majeurs de séquençage des génomes. Puisqu'il était connu que seulement une faible fraction du génome des animaux codait pour des gènes, il était plus économique et efficace de séquencer des

ARNm extraits pour trouver des gènes que de s'intéresser au génome complet (Waterston et al., 1992). Chaque nouvelle technologie apporte son lot d'adjectifs forts pour décrire son apport à la science. Les projets d'EST sont décrits comme produisant de larges quantités de séquences. Il est important de relativiser avec l'époque pré-HGP, où la première base de données d'EST possédait moins de 15 000 séquences d'*Homo sapiens* à sa publication originale, en 1993 (Boguski, Lowe, & Tolstoshev, 1993).

Séquençage d'ADNc. Avec le développement des technologies de séquençage plus efficaces et moins coûteuses, le séquençage d'EST a progressivement laissé sa place au séquençage des données d'annotation appelée ADNc. Le terme ADNc est imprécis, puisqu'il décrit un type de molécule, issue de la transcription inverse de l'ARN, technique qui était aussi utilisée en production d'EST. On peut différencier la production d'EST et d'ADNc comme étant le séquençage d'ARN rétrotranscrits respectivement non-fragmentés et fragmentés. Alors que les premières données d'EST ont été générées par séquençage de type Sanger, soit le séquençage de première génération, le séquençage massif d'ADNc est issu du développement du séquençage de seconde génération (NGS, Next-Generation Sequencing). Le NGS permet d'obtenir un nombre beaucoup plus élevé de lectures que le séquençage de type Sanger, mais en échange de lectures de plus petite longueur. Avec ce changement technologique, il était possible de s'intéresser à l'entièreté de l'ADNc, en le fragmentant et en le séquençant, afin d'obtenir des lectures uniformément distribuées le long des ARNm originaux. Cette technique permet donc, en plus d'identifier les frontières en amont et aval des gènes, d'observer l'épissage alternatif des ARNm.

1.3.2 Développement en cours

L'évolution des annotations suit le développement technologique, et les prototypes de nouvelles technologies de séquençage présents sur le marché annoncent les prochains développements potentiels de données utiles à la définition des annotations.

ADNc pleine longueur. La fragmentation des ARN est une nécessité issue d'une limitation technologique du séquençage de seconde génération. Une des prochaines étapes pour l'obtention de données plus fidèles et fiables à la création d'annotations serait la capacité de séquencer des ADNc pleine longueur. Admettant un processus d'extraction de l'ARN qui ne produit pas d'altération, le séquençage pleine longueur permettrait d'observer directement la population d'ARN possible, ainsi que les différents isoformes formés de

combinaison d'exons qui peuvent être très difficile à résoudre par NGS. Certains projets utilisant cette technologie commencent à émerger. L'analyse du transcriptome du porc par NGS et Iso-seq (séquençage d'isoforme), soit une technologie de séquençage d'ADNc de pleine longueur, a permis la validation des modèles de gènes actuels, mais aussi la découverte de nouveaux isoformes de gènes connus, et de nouveaux gènes, en comparaison aux annotations d'Ensembl de la même espèce (Beiki et al., 2019).

Séquençage d'ARN direct. Les annotations génomiques s'intéressent à détailler les produits possibles de la transcription. Tous les types de données précédemment énoncés sont en fait séquencés sous forme d'ADNc, issu de la transcription inverse de l'ARN. Cela signifie que les annotations génomiques sont définies avec l'hypothèse d'une transcription inverse non biaisée. Des travaux récents de notre laboratoire ont démontré que l'utilisation d'une transcriptase inverse opérant à plus haute température que l'enzyme usuelle produit un transcriptome avec des proportions de transcrits par biotype plus fidèles aux autres méthodes quantitatives biochimiques (Boivin et al., 2018). Ainsi, le développement de méthode de séquençage d'ARN direct, sans étape de transcription inverse, permettrait potentiellement de contourner le biais de la transcriptase inverse (Workman et al., 2019). Il reste aussi encore à démontrer que le biais observé chez les transcriptases, où l'hypothèse est que les ARN hautement structurés ne peuvent qu'être inversement transcrits à haute température, ne se propage pas aussi chez les porines présentement utilisées dans le séquençage direct d'ARN.

1.4 Annotations génomiques chez *Homo sapiens*

Il existe, et a existé, plusieurs annotations génomiques concurrentielles chez *Homo sapiens*. Une liste non exhaustive de ces annotations est présentée dans cette section. Le domaine de la bio-informatique est caractérisé par la multiplicité de ses ressources (Ison et al., 2016), ce qui rend difficile n'importe quel effort de description intégral des choix disponibles. On peut aussi se questionner sur la pertinence de cette multiplicité, et sur l'utilisation de chacune des ressources dans la communauté scientifique.

Retracer l'histoire de l'évolution des projets d'annotations génomiques n'est pas une tâche triviale, puisque cette histoire est truffée de projets indépendants, de fusions partielles, d'abandon de projets, et ce, pas toujours répertorié de façon explicite.

1.4.1 Projets d'annotation

HAVANA. HAVANA est un projet majeur d'annotation manuelle, développé par le *Wellcome Sanger Institute* (Wellcome Sanger Institute, 2019). Son objectif originel est d'annoter tous les transcrits possibles chez l'humain, la souris, le poisson-zèbre et le rat. Le projet s'intéresse plus récemment à l'annotation des pseudogènes et des transcrits avec de l'épissage alternatif complexe, car ce sont des cas plus difficiles à résoudre pour les méthodes d'annotation automatique. Le projet HAVANA utilise des données de séquences protéiques, d'EST et d'ADNc pour générer ses annotations. Les limitations d'HAVANA sont celles issues d'un travail qui repose sur les décisions d'individus. La production d'annotation manuelle est coûteuse, lente, difficile à mettre à l'échelle et donc à suivre l'abondance de nouveaux génomes et nouvelles espèces séquencées, et surtout peu agiles, où si un changement majeur de méthode d'annotation ou d'hypothèse biologique est effectué, il faudra revenir manuellement sur les cas problèmes. Il faut aussi s'assurer de la concordance et de l'uniformité du travail accompli par les personnes qui assurent la curation des données (Loveland, Gilbert, Griffiths, & Harrow, 2012).

Les annotations manuelles d'HAVANA sont distribuées sur la plateforme VEGA (*Vertebrate and Genome Annotation*), qui est maintenue à jour par l'EBI (*European Bioinformatics Institute*) (Wilming et al., 2008). Par contre, la dernière mise à jour de VEGA date de février 2017 et VEGA a annoncé un maintien de la base de données sous forme d'archive pour trois ans, soit jusqu'en février 2020. Après cette date, les données d'HAVANA ne sont qu'incluses dans la distribution d'Ensembl et de GENCODE.

Ensembl. Ensembl est la ressource d'annotations génomiques distribuée par l'EBI. Les annotations génomiques d'Ensembl sont issues de la combinaison des données manuelles d'HAVANA, et du pipeline de prédiction d'Ensembl appelé Ensembl-genebuild (Frankish et al., 2019). Chaque entrée de l'annotation d'Ensembl est affiliée à la méthode qui l'a créé, soit manuelle ou automatique.

Comme point de départ de prédiction, Ensembl utilise les données d'HAVANA et des données de prédiction de structure de gène qui n'utilise que la séquence génomique (Burge & Karlin, 1997). Suivant cela, les séquences de protéines, d'ADNc et d'EST sont alignées sur le génome et puis combinées pour identifier les séquences codantes et les UTR (Untranslated region). Les gènes sont ensuite formés des différents transcrits qui partagent

au moins un exon. Les données de gènes et de protéines d'autres espèces peuvent être utilisées, par étude d'homologie, pour aider la prédiction (Curwen et al., 2004).

Il est intéressant de noter que les deux parties d'Ensembl, Ensembl-HAVANA et Ensembl-genebuild, interagissent en rétroaction, les deux pouvant être le point de départ de l'autre. Les annotations automatiques donnent un point de départ aux annotations manuelles, et les annotations manuelles peuvent être utilisées pour valider et informer les méthodes automatiques.

GENCODE. ENCODE (*Encyclopedia Of DNA Elements*) est un consortium public de recherche dont l'objectif est l'étude des divers éléments fonctionnels du génome humain. GENCODE en est un sous-projet qui s'intéresse à l'annotation des gènes et de leurs produits. Les ressources publiques de GENCODE distribuent des annotations génomiques pour l'humain et la souris. Pour ces deux espèces, les annotations de GENCODE sont dorénavant identiques à celles distribuées par Ensembl, sauf exempt d'un léger décalage temporel possible. Historiquement, GENCODE représentait la fusion d'HAVANA et d'Ensembl-genebuild, mais depuis qu'HAVANA a intégré le projet Ensembl, les deux annotations sont correspondantes. À noter que GENCODE et Ensembl utilisent des systèmes de numérotation différents pour indiquer les versions de leurs annotations respectives.

RefSeq. RefSeq (*Reference Sequence*) est la ressource d'annotations génomiques développée par le NIH (*National Institutes of Health*). RefSeq a été créé avec la vision de définir une base de données centrale permettant l'identification non ambigüe des gènes, de leurs produits et leurs séquences appropriées (Maglott, Katz, Sicotte, & Pruitt, 2000). L'idée était d'éviter la duplication des ressources, ce qui mène à la création de plusieurs identifiants pour les mêmes gènes et produits.

RefSeq met beaucoup d'importance sur les données issues de la curation manuelle. Les informations issues de la curation manuelle peuvent être soumises par la communauté scientifique, entrent dans la base de données sous statut provisoire, puis se voient réviser par la curation interne de RefSeq. Les différents gènes et leurs produits issus de la curation manuelle sont appuyés par des publications scientifiques les ayant étudiés.

RefSeq possède aussi deux pipelines d'annotations automatiques, soit un pour eucaryote et un pour procaryote (O'Leary et al., 2016). Le pipeline eucaryote ressemble beaucoup

à celui d'Ensembl-genebuild, soit la prédiction de gènes à partir de l'ADN, appuyé par les annotations manuelles, puis ajout d'information de séquence de protéines, d'ADNc, et d'EST (Thibaud-Nissen, Souvorov, Murphy, DiCuccio, & Kitts, 2013). Par contre, les logiciels utilisés dans les deux pipelines sont différents les uns des autres, et donc on peut s'attendre à des annotations non identiques.

UCSC. En génomique, UCSC (*University of California Santa Cruz*) est surtout connu pour le navigateur de génome, nommé le UCSC *Genome Browser* (Kent et al., 2002). Cet outil iconique a comme objectif de procurer un visuel aux différentes informations d'annotation sur le génome. Ces informations inclues les annotations génomiques, soit la position et donc la séquence des gènes et de leurs produits, mais aussi d'autres informations comme la présence de motifs connus dans l'ADN, ou bien des informations de conservation à travers les espèces, ou encore des EST et des ADNc alignés sur le génome. Ainsi, le UCSC Genome Browser a été pendant plusieurs années une ressource où il était possible de consulter les annotations des différents projets existants.

En 2003, UCSC a aussi annoncé sa propre annotation génomique sous le nom de UCSC *Known Genes*, qui a été publiée en 2005 (Hsu et al., 2006). UCSC avait le désir de fournir une annotation beaucoup plus agilement que RefSeq, principale annotation disponible sur le UCSC *Genome Browser* à l'époque, en se basant quasi exclusivement sur une méthode d'annotation automatique. UCSC s'appuyait sur les données protéiques d'UniProt pour effectuer ses prédictions (Bairoch et al., 2005), tout en se limitant à la prédiction des gènes codants et des ARN messager.

En juillet 2015, UCSC annonçait, par le biais de son blogue, que le UCSC *Genome Browser* allait changer de source pour l'annotation de base, passant de UCSC *Known Genes* à GENCODE (UCSC Genome Browser, 2015). Par le fait même, UCSC a arrêté le développement de son annotation pour l'humain, et ne distribue dorénavant que des tables permettant de convertir les identifiants de transcrits d'UCSC vers les identifiants de GENCODE. Dans les motivations pour l'arrêt du support des UCSC *Known Genes*, il a été mentionné qu'il est bénéfique de réduire le nombre d'annotations compétitrices pour la communauté bio-informatique, afin de réduire le potentiel de confusion entre les différentes ressources.

Autres projets d'annotation. Il existe d'autres projets d'annotation chez l'humain qui seront mentionnés dans ce mémoire, mais qui ne seront pas abordés en détail ici. Certains projets, comme AceView (Thierry-Mieg & Thierry-Mieg, 2006) et le H-invDB (Yamasaki et al., 2009) n'ont pas été mis à jour depuis des années, et d'autres, comme CHESS (Pertea et al., 2018) sont très récents et fort peu utilisés dans la littérature.

1.4.2 Qualité des annotations génomiques

La qualité et la concordance des annotations génomiques sont caractérisées par plusieurs métriques internes ou comparatives.

Ensembl. Ensembl utilise les TSL (*Transcript Support Level*) pour identifier la qualité de l'annotation des transcrits. Les TSL se déclinent en cinq niveaux de qualité d'annotation, où chaque niveau correspond aux preuves utilisées pour établir ladite annotation. Pour être TSL1, soit la meilleure qualité possible, toutes les jonctions d'épissage d'un transcrit doivent être supportées par au moins un ADNc fiable. TSL3, qualité médiane, indique le même type de support, mais par des EST. À TSL5, pire niveau de qualité, il y a des jonctions d'épissage du modèle qui ne sont pas supportées par des données biologiques. Puisque les TSL sont définis en fonction du support des jonctions d'épissage, les transcrits mono-exoniques, bien souvent des ARN non codants, ne peuvent avoir de TSL, et sont donc identifiés par la mention TSLNA. De manière intéressante, le système TSL ne distingue pas en fonction de la nature de l'annotation. Les mêmes critères sont utilisés pour une annotation manuelle ou automatique, seule la nature des preuves est regardée.

Ensembl possède aussi les *Golden Transcripts*, ou transcrits dorés, qui sont des transcrits qui ont été trouvés de manière identique par Ensembl-HAVANA et Ensembl-genebuild (Ensembl, 2020).

RefSeq. Le seul indice de qualité contenu dans RefSeq est la première lettre de l'identifiant des gènes, transcrits et protéines, soit N ou X, qui veulent respectivement dire que l'objet identifié est issu de curation manuelle ou de prédiction automatique.

HGNC. Si l'on veut pouvoir comparer des annotations génomiques issues de différents projets, il est nécessaire d'avoir une base de comparaison. Bien que cette problématique

semble triviale, les gènes ne possèdent pas nécessairement les mêmes noms à travers les différentes ressources, et ne sont pas nécessairement reliés en relation un pour un. Pour faire le pont entre les gènes, le HGNC (*HUGO Gene Nomenclature Committee*) a mis en place un identifiant commun pour les gènes humains (Yates et al., 2017). En plus de l'identifiant unique (de format numérique), le HGNC publie le symbole officiel des gènes humains. Le symbole est un identifiant alphanumérique souvent utilisé comme le nom usuel prononçable des gènes. Par exemple, SRSF1 est le symbole officiel d'un gène aussi connu comme HGNC :10780, ENSG00000136450 et GeneID :6426 chez respectivement HGNC, Ensembl et RefSeq. SRSF1 possède aussi des symboles alternatifs, comme SF2, qui sont utilisés dans la littérature, mais le symbole officiel distribué par HGNC est à favoriser pour l'uniformité de la littérature. Les données incluses dans le HGNC proviennent à la fois de la curation interne, et des données proposées par les autres projets d'annotation, ce qui en fait un projet collaboratif. Par contre, le projet se limite à faire le pont entre les gènes, sans s'avancer dans la correspondance entre transcrits.

APPRIS. APPRIS (*Annotation of principal and alternative splice isoforms*) est un projet développé par un groupe de recherche indépendant, membre du consortium GENCODE. Le projet APPRIS s'intéresse à l'annotation fonctionnelle des gènes, et surtout de la relation fonctionnelle entre les différents isoformes d'un même gène. Pour ce faire, APPRIS défini un isoforme comme étant l'isoforme principal du gène, soit l'isoforme représentant la fonction cellulaire principale du gène (Rodriguez et al., 2013). L'hypothèse ici étant qu'un gène possède une fonction première, et que l'épissage alternatif vient altérer cette fonction dans certains cas de figure. Les deux sources de données utilisées pour déterminer l'isoforme principal sont l'expression à travers les différents tissus, où l'isoforme principal devrait être exprimé le plus ubiquitairement que les autres isoformes, et la conservation, où l'isoforme principal devrait être le plus conservé à travers les espèces. Le projet APPRIS ne s'intéresse présentement qu'aux gènes codants, où l'isoforme principal est nécessairement un isoforme produisant une protéine.

CCDS. Le CCDS (*Consensus CDS*) est un projet collaboratif, publié en 2009, de convergence des CDS chez les trois grandes annotations génomiques de l'époque, soit Ensembl, RefSeq et UCSC, pour l'humain et la souris (Pruitt et al., 2009). L'idée du projet CCDS est de déterminer les CDS qui sont identiques à travers les différentes annotations du projet, de les identifier avec un numéro d'accession unique, et d'en conserver une liste qui sera considérée comme porteuse des séquences les plus fiables. Le projet se veut aussi une

plateforme encourageant la curation manuelle entre les bases de données d'annotation afin de converger les CDS similaires, mais non identiques des mêmes gènes. L'information considérée par le CCDS n'est que la séquence codante, excluant toute information des UTR et des ARN non codants. Bien que UCSC ne soit plus, le développe du projet CCDS continue avec Ensembl et RefSeq.

1.5 Utilisation des annotations en RNA-seq

Dès lors qu'une analyse bio-informatique rapporte de l'information en fonction de gènes, de transcrits ou de protéines, il y a nécessairement une annotation génomique d'impliquée, puisque ces objets biologiques ont besoin d'être définis quelque part afin d'être étudiés. Comme énoncé précédemment, la combinaison de génome de référence et d'annotation génomique est la source première de séquence biologique.

Le séquençage de l'ARN, aussi nommé RNA-seq, est utilisé pour identifier et quantifier relativement les transcrits présents dans un échantillon biologique. L'information contenue dans l'annotation génomique peut être intégrée dans le pipeline de RNA-seq à plusieurs étapes. Après l'obtention des lectures par séquençage et la vérification de leur qualité, les lectures doivent être alignées sur une référence biologique, afin d'identifier de quels gènes elles ont été produites. Les deux choix de référence sont le génome, soit l'entièreté des séquences d'ADN connues de l'espèce étudiée, ou bien le transcriptome, soit l'entièreté des séquences du génome connues comme pouvant être transcrites. Pour générer un transcriptome, il faut extraire toutes les séquences des transcrits matures contenus dans une annotation génomique, en ne conservant que les séquences contenues dans les exons. Outre la quantité différente d'information contenue dans le génome et dans le transcriptome, la principale différence entre ces deux méthodes est l'absence d'intron dans le transcriptome (Yi, Liu, Melsted, & Pachter, 2018). Cela veut dire que, pour aligner les lectures sur le génome, le logiciel doit pouvoir fragmenter des lectures pour les aligner de façon non contigu à travers les jonctions d'épissage. Dans ce cas d'alignement génomique, certains logiciels d'alignement peuvent recevoir une annotation génomique en option, et l'utilisent pour aider la résolution des jonctions d'épissage. Les annotations génomiques peuvent aussi être incluses à l'étape de quantification, où elles sont utilisées pour informer les coordonnées génomiques où les lectures ont été alignées. Ainsi, il est possible de rapporter la quantification en fonction des différents niveaux hiérarchiques des annotations, soit les gènes, transcrits ou exons. Ainsi, que ce soit pour générer un transcriptome, aider

l'alignement non contigu ou servir de base de quantification, une expérience de RNA-seq se doit d'utiliser une annotation génomique (Conesa et al., 2016).

1.6 Impact du choix d'annotation génomique en RNA-seq

L'existence de plusieurs références amène la nécessité de sélectionner une de ces références lors d'une étude, ou de comparer les différents résultats possibles. Cette problématique, déjà illustrée par les projets comme HGNC et CCDS, est connue et a déjà fait l'objet de plusieurs publications scientifiques. Cette section détaille différents travaux s'intéressant à la question de l'impact des annotations en RNA-seq.

Une des problématiques dans l'établissement de métriques d'étalonnage est de s'assurer qu'elles démontrent bien ce qu'on essaie de mettre de l'avant. La problématique d'identification de méthode optimale en RNA-seq est complexifiée par l'absence d'étalon de référence, d'expérience théorique parfaite qui pourrait être utilisée pour identifier l'erreur absolue des méthodes étudiées.

Wu , Phan et Wang (2013) ont étudié l'impact du choix d'annotation génomique sur des métriques d'alignement, en relation avec ce qu'ils décrivent comme la complexité des annotations. Les auteurs décrivent la complexité comme étant proportionnelle au nombre de gènes présents dans l'annotation. Les six annotations considérées, de la plus complexe à la moins complexe, sont AceView, H-InvDB, Ensembl, VEGA, UCSC et RefSeq. L'ordre des annotations reste le même si on considère le nombre total de transcrits, le nombre total d'exons ou le nombre de transcrits par gène comme métrique de complexité. L'article note que plus les annotations sont complexes, plus il y a de lectures qui sont multi alignées, c'est-à-dire qui peuvent être alignées à plusieurs endroits avec la même qualité d'alignement. Aussi, plus l'annotation est complexe, plus les lectures ont de chance d'être alignées sur des régions annotées. Il a aussi été observé que moins l'annotation est complexe, plus elle a de gènes exprimés. Pour évaluer l'impact des annotations sur la quantification, le coefficient de variation moyen de l'expression des gènes à travers plusieurs réplicats a été utilisé. Pour les gènes en commun entre les différentes annotations, il n'y avait pas de différence notable, alors que pour les autres gènes, plus l'annotation était complexe, plus la quantification était variable. Les auteurs concluent avec l'idée d'utiliser la complexité pour définir l'utilisation potentielle d'une annotation, où moins complexe veut dire quantification plus fiable, et plus complexe veut dire potentiel de découverte plus large.

Frankish et als. (2015) se sont intéressés à décrire les différences d'annotation entre GENCODE (Ensembl) et RefSeq. Il y est noté que, bien que RefSeq est souvent considéré dans la mémoire populaire comme étant une annotation plus fiable, puisque plus manuelle, 93.4% des transcrits de gènes codants de GENCODE v21 étaient manuellement curés, ou identiques dans la curation manuelle et la prédiction, alors que RefSeq, à la même époque, était à environ 51%. Les auteurs avancent aussi que le pipeline de RefSeq est beaucoup plus contraignant dans l'utilisation des mêmes exons initiaux et finaux pour tous les transcrits d'un gène, n'admettant pas l'utilisation large de sites alternatifs d'initiation et de terminaison de la transcription. Il est avancé dans la littérature que les sites alternatifs d'initiation et de terminaison de la transcription seraient responsables de plus de variabilité dans la définition des transcrits d'un même gène que l'épissage alternatif (Reyes & Huber, 2018). Outre comparer les différences dans le type d'évènement d'épissage alternatif entre RefSeq et GENCODE, les auteurs avancent aussi la même idée de complexité que Wu, Phan et Wang (2013), où l'utilisation d'une annotation moins complexe serait plus apte à la quantification fiable, et l'utilisation d'une annotation plus complexe serait plus apte à la découverte.

Avec une approche plus théorique, Pyrkosz, Cheng et Brown (2013) ont exploré, par simulation, l'impact de la complétude d'une annotation génomique sur la qualité de la quantification obtenue. Ils concluent que l'absence de transcrits, qui sont biologiquement exprimés, de l'annotation peut affecter largement d'autres gènes, en produisant des alignements de qualité inférieurs dans plusieurs autres sites, au lieu de l'alignement optimal dans le site manquant. Mais ils avancent aussi l'inverse, que la présence dans l'annotation génomique de trop de transcrits non exprimés dans l'échantillon biologique peut aussi avoir un impact négatif sur la bonne quantification. On se retrouve donc dans une problématique assez difficile à définir. La solution avancée par Pyrkosz, Cheng et Brown serait l'utilisation de lectures de séquençage assez longues, plus de 2000 nucléotides selon eux, pour pouvoir mieux identifier ce qui est réellement transcrit.

Ces articles sont un sous-ensemble non exhaustif de la littérature, mais ils représentent bien l'état de la caractérisation des annotations. Les analyses disponibles regardent beaucoup le problème de loin, en se cachant derrière des métriques très peu informatives, comme le nombre de gènes et les pourcentages d'alignement (Ballouz et al., 2018). Frankish et als. (2015) décrivent des différences dans la structure des annotations entre Ensembl et RefSeq, mais ne les relient pas avec des biais de résultats. Les biais potentiels ont aussi peu de chance d'être uniformément distribués à travers les gènes, mais peu d'analyses

s'attardent à décrire la problématique à un niveau plus bas. Il n'y a aussi aucune étude, à notre connaissance, qui insère la problématique de biais des annotations dans le contexte du RNA-seq, soit en comparaison aux autres biais existants et décrits.

1.7 Problématiques et objectifs

Les annotations génomiques sont des ressources centrales qui regroupent nos connaissances sur la nature des gènes et qui servent de base à plusieurs types d'analyses bio-informatiques. La multiplicité des annotations démontre qu'il n'y a pas encore de vision uniforme sur la biologie et la structure des gènes et de leurs produits.

Cette multiplicité crée aussi un besoin de choix de données de référence lors d'analyse bio-informatique. Ce choix se doit d'être informé, mais il s'avère difficile de démontrer la supériorité d'un choix d'une annotation sur les autres pour une analyse donnée. Il semble y avoir beaucoup de problématiques différentes convoluées qui ont un impact, dont le choix des divers logiciels dans le pipeline d'analyse, les gènes étudiés, ce qui inclut la nature de leurs transcrits, de leur mécanisme d'épissage et de sites de transcription alternatifs et autres propriétés issues du positionnement dans le génome.

Ainsi, au lieu de travailler sur la description de la méthode optimale de RNA-seq à utiliser, ce mémoire s'intéresse plutôt à détailler la place des annotations génomiques dans le RNA-seq et le biais potentiel qu'elles créent. Il est déjà connu que le pipeline bio-informatique de RNA-seq n'est pas parfait, et que les résultats vont varier en fonction des autres choix expérimentaux que celui de l'annotation génomique (Teng et al., 2016 ; Sahraeian et al., 2017 ; Williams, Baccarella, Parrish, & Kim, 2017). La question importante est donc : quelle est l'importance du biais apporté par les annotations génomiques sur le pipeline de séquençage de l'ARN, en relation avec le biais des autres étapes ? Ainsi, cela permettra de mieux comprendre l'importance de la problématique, et la potentielle urgence à l'encadrer.

Le projet se décline donc en deux parties principales, qui seront détaillées dans les deux prochains chapitres :

– Détailler les choix méthodologiques utilisés en RNA-seq dans la littérature,
– Étudier et déconvoluer les biais principaux du pipeline de quantification de RNA-seq.

Chapitre 2
Article 1 : Current RNA-seq methodology reporting limits reproducibility

Current RNA-seq methodology reporting limits reproducibility

Auteurs de l'article: Joël Simoneau, Simon Dumontier, Ryan Gosselin, Michelle S. Scott

Statut de l'article: Article publié dans *Briefings in Bioinformatics*. 2019 Dec 10 ; [DOI: 10.1093/bib/bbz124]

Avant-propos: L'objectif du premier article est la caractérisation de la méthodologie utilisée dans le pipeline computationnel en séquençage de l'ARN. Bien qu'il existe une vaste gamme d'outils et de références biologiques, un petit nombre d'outils occupe la majeure partie des méthodologies utilisées. L'article met aussi en lumière que certaines informations, principalement l'annotation génomique utilisée, sont souvent ignorées et non rapportées dans les méthodologies. À notre connaissance, cette étude est la première à caractériser largement les pipelines utilisés en RNA-seq dans la littérature, et elle ouvre la voie à une étude de l'impact méthodologique des principaux outils utilisés.

Contributions : JS et SD ont effectué la curation des données. JS a effectué l'analyse des données. JS, RG et MSS ont conceptualisé la recherche et participé à la rédaction du manuscrit.

Résumé

Le séquençage des acides ribonucléiques (RNA-seq) identifie et quantifie les molécules d'ARN d'un échantillon biologique. La transformation des données brutes de séquençage pour l'obtention de données de quantification de gènes ou d'isoformes requiert un pipeline bio-informatique. Ces pipelines sont modulaires, construits à partir d'un ensemble de logiciels et de données de référence biologique. Les logiciels sont habituellement choisis et paramétrés en fonction du protocole de séquençage choisi et des questions biologiques. Par contre, alors que le bruit technique et biologique est réduit par l'utilisation de réplicats, les biais dus au choix de pipeline et de références biologiques sont souvent négligés. Dans cette étude, nous démontrons que la pratique courante de publication empêche la reproductibilité des études de RNA-seq en ne spécifiant pas les informations méthodologiques pertinentes. Les articles scientifiques avec comité de lecture sont censés appliquer les méthodologies standards et scientifiquement approuvées. Dans la mesure où le pipeline de RNA-seq optimal et sans biais n'est pas parfaitement défini, toute information méthodologique possède un rôle important dans la définition des résultats d'une étude. Ce travail illustre la nécessité d'avoir une méthode explicite et standardisée pour la communication des informations méthodologiques pour les expériences de RNA-seq.

2.1 Abstract

Ribonucleic acid sequencing (RNA-seq) identifies and quantifies RNA molecules from a biological sample. Transformation from raw sequencing data to meaningful gene or isoform counts requires an in silico bioinformatics pipeline. Such pipelines are modular in nature, built using selected software and biological references. Software is usually chosen and parameterized according to the sequencing protocol and biological question. However, while biological and technical noise is alleviated through replicates, biases due to the pipeline and choice of biological references are often overlooked. Here, we show that the current standard practice prevents reproducibility in RNA-seq studies by failing to specify required methodological information. Peer-reviewed articles are intended to apply currently accepted scientific and methodological standards. Inasmuch as the bias-less and optimal RNA-seq pipeline is not perfectly defined, methodological information holds a meaningful role in defining the results. This work illustrates the need for a standardized and explicit display of methodological information in RNA-seq experiments.

2.2 Introduction

Ribonucleic acid sequencing (RNA-seq) enables the identification and quantification of RNA molecules from a biological sample. Microarrays, long considered the state of the art for large-scale RNA quantification, need a known genome annotation for probe design, prior to the actual experiment [1]. In contrast, in an RNA-seq experiment, a genome annotation is introduced after the sequencing step, permitting one to reanalyze the same dataset using different software and references [2], maintaining the relevance of datasets after genome and genomic annotation updates. In the past decade, RNA-seq has been rapidly democratized due to a dramatic lowering of the cost of sequencing. This has led to the diversification and multiplication of sequencing analysis applications, and the generation of large numbers of datasets to analyze. Hundreds of different software applications have been published to fulfill this need in a very modular fashion [3], allowing the usage of custom in silico pipelines defined by user-selected software and references for each analytical step.

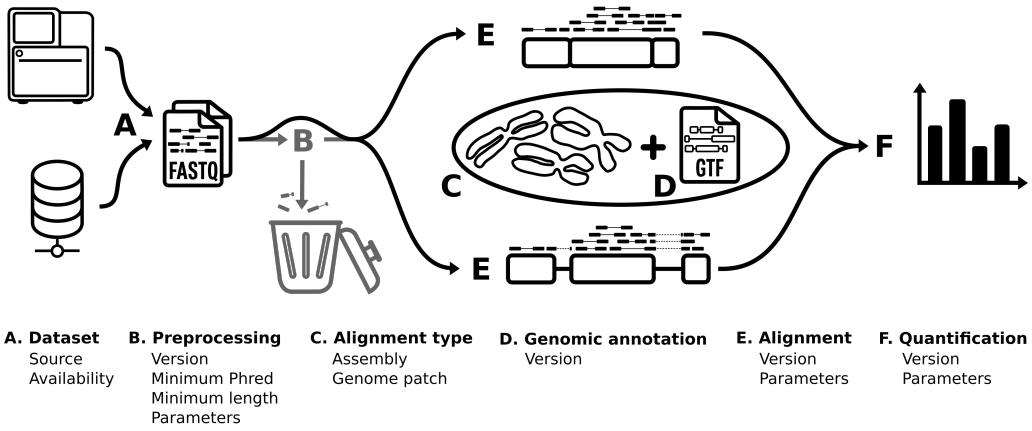


Figure 1 – RNA-seq bioinformatics pipeline. Schematic of the RNA-seq bioinformatics methodology. The pipeline is divided into six steps (A–F). Each step is specified using a series of parameters displayed on the figure.

Quantification results in RNA-seq are subject to different types of noise which are usually categorized as either technical or biological in nature [4]. Technical noise represents variations due to the laboratory manipulations, from RNA extraction to sequencing. Biological noise includes a broader array of sources, depending on the experimental design. This covers differences that can go from non-genetic individuality of cells exposed to homogeneous conditions [5], all the way to inter-individual genetic variations. Biological and technical sources of noise are stochastic in nature, and thus perfect reproducibility is impossible to achieve. Replicates can be used to quantify the variations and alleviate their impacts [6]. Several studies have already investigated the number of replicates needed for reproducible results [7–9]. However, another source of discrepancy seems to have often been overlooked. The optimal RNA-seq in silico processing pipeline, from raw sequencing files to meaningful gene or isoform counts, has not yet been (and perhaps will never be) defined. Thus, the same data can be processed in a multitude of ways, using different combinations of modular software and references. The distribution of a gene's quantification across all the different RNA-seq pipelines (used or possible) is what we have dubbed the 'in silico design noise'. Theoretically, such noise is deterministic in nature. While some software does have stochastic processes, these are undesirable when they cause random variability in the results. Given the exact same inputs (i.e. same dataset, software and parameters), it is preferable to always obtain the same results. Therefore, this 'in silico design noise' is actually a function of the software and parameter selection. One can reduce the impact of such noise and ensure the reproducibility of the analysis by explicitly specifying every information related to data transformations used to process the RNA-seq datasets. In this survey, we scrutinize the exhaustivity of the scientific literature regarding reported

methodology of RNA-seq experiments. We find that only a minority of articles (25%) describe all essential computational steps and fewer still specify all parameter values to ensure complete reproducibility. From these analyses, we stress that a better disclosure of methodological information by users, developers and editors will beget more reproducible scientific literature.

2.3 Discussion

2.3.1 Standard steps in an RNA-seq computational analysis pipeline

While RNA-seq experiments can be analyzed in different ways, in organisms with annotated genomes, RNA-seq computational pipelines typically all follow the same series of steps to obtain a quantification of transcripts from a raw read file (Figure 1), following which, diverse further analyses are possible [10]. Here, we focus on the first steps that are common to all RNA-seq pipelines, from read file to transcript quantification, consisting of several design choices of tools and references that are essential to use in order to compute quantification and to specify in order to ensure reproducibility of the results (Figure 1). The first such element is the description of the source of the raw read file, whether the RNA-seq data were generated in the context of the current study or whether the data were obtained from a repository such as the Gene Expression Omnibus. It is common to preprocess the data by verifying their quality, trimming reads of lower quality and removing non-biological sequences (i.e. sequencing adapters and indexes) before performing the alignment. We note however that current quality values per nucleotide are such that a preprocessing step might increasingly be considered unnecessary. While every other step is essential in RNA-seq, preprocessing is the only step that produces the same file type as it uses. This means that it does not, in a computational manner, need to be run. The alignment type (whether reads are aligned on the genome or the transcriptome) and the annotation file used to define the genomic features considered are also essential elements of methodological information. Finally, once the source of the reads and the identity of the genome/transcriptome are defined, one must also indicate the tools used to align the reads to the genome/transcriptome and to quantify the abundance of the different genomic features annotated. Each of these steps requires specification of parameter values to entirely describe how the step was carried out (Figure 1).

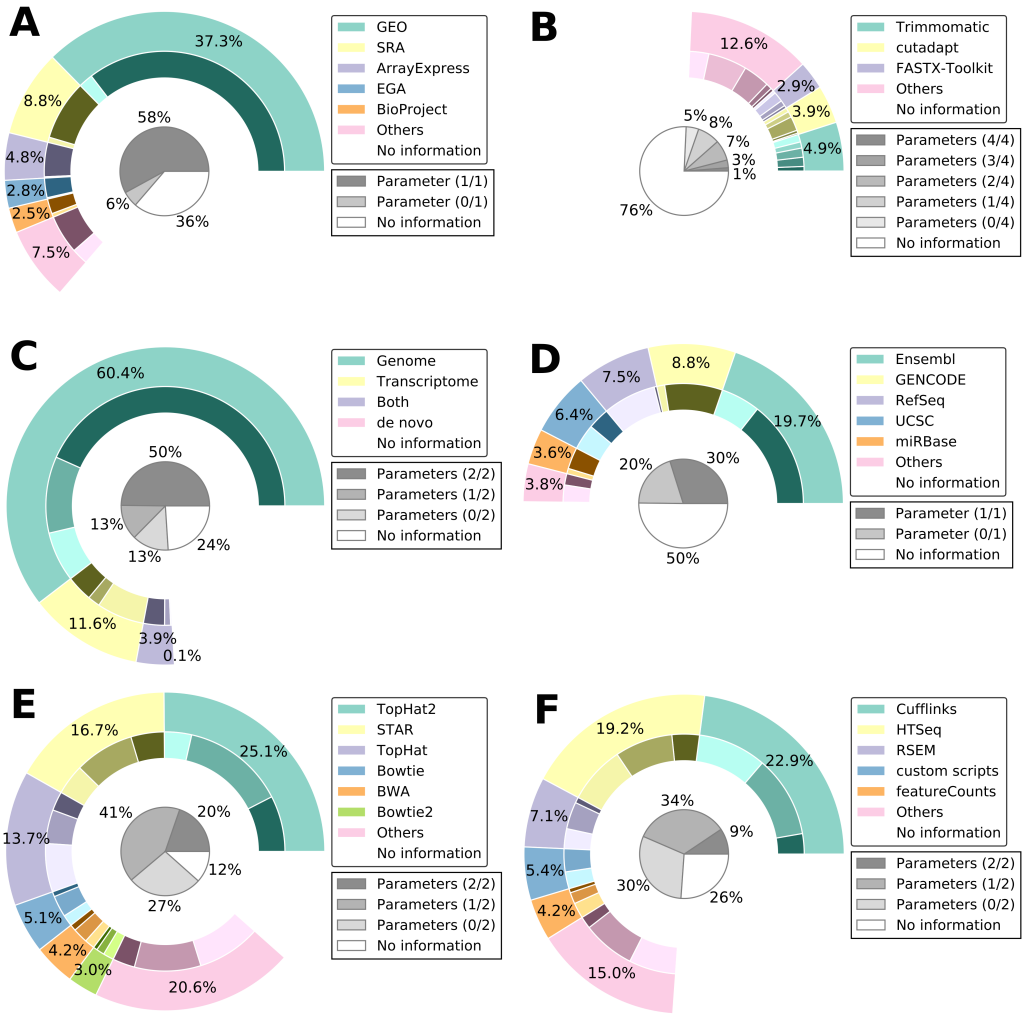


Figure 2 – RNA-seq reported methodology is incomplete. Distribution of software and reference usage for the six methodological steps of an RNA-seq experiment (**A.** dataset, **B.** preprocessing tool, **C.** alignment type, **D.** genomic annotation, **E.** alignment tool and **F.** quantification tool). The outer donut chart illustrates the distribution of the primary criterion for each step. The inner donut chart illustrates the degree of parameter specification : the darker the shade, the more complete the information. The inner pie chart is the summation of all shades from the inner donut. Complete results are available as Supplementary data.

2.3.2 RNA-seq pipelines are diverse, consisting of many different software tools and references

To investigate methodology reporting practices in RNA-seq computational analysis pipelines, 1000 randomly chosen articles performing RNA-seq were analyzed by two independent reviewers as described in the Methods (Supplementary data). To ensure only consideration of articles with comparable pipelines, we kept only articles that included Methods starting from a raw read file and obtaining transcript or gene quantification. A single species study was necessary due to species-specific references used in RNA-seq.

Homo sapiens was chosen due to the large availability of studies and references. The article exclusion criteria are described in Supplementary Table 1 and following their application, 465 articles remained. From these articles, information about the RNA-seq pipeline, from FASTQ files to gene or isoform count matrices, was extracted to determine which tools and parameters are used in the RNA-seq literature.

As shown in Figure 2, many references and tools are commonly used in RNA-seq computational pipelines. Most RNA-seq analyses align reads against a genome rather than a transcriptome (Figure 2C), but the choice of reference annotation is distributed among several sources (Figure 2D). For the three steps involving tool choices (Figure 2 panels B, E and F representing the preprocessing, alignment and quantification steps, respectively), no tool is reported to be used in more than 25% of the articles and thus several different tools are common in RNA-seq pipelines. We note however that most commonly used alignment tools are part of the Tuxedo suite.

RNA-seq is a young and quickly evolving field. The commonly used sequencing parameters, in terms of read length, depth and read pairing, have changed in the past years, and new software is being developed to better use this more informative data [10, 11]. But we observed latency in tool usage in the literature. While older tools need to be readily available for comparative studies, users generating new results should be encouraged to modernize their pipelines and move to updated or newer tools. We have highlighted the use of the Tuxedo suite family (TopHat, TopHat2, HISAT and HISAT2) [12–14] to observe that TopHat, while being three times reimplemented (Figure 3), is still being used as an aligner for newly published studies despite indications in the last TopHat2 release that HISAT2 should now be preferred (in TopHat 2.1.1 release 23 February 2016). Ever increasing numbers of computational resources are becoming available, and newer software is usually more computationally efficient. Efforts to reanalyze studies with state-of-the-art pipelines should be put forward. The fast pace nature of RNA-seq technology requires one to be agile for timely contributions. If updating software to newer, benchmarked and better-performing software has an impact on the results, then one can argue that the implementation was useful.

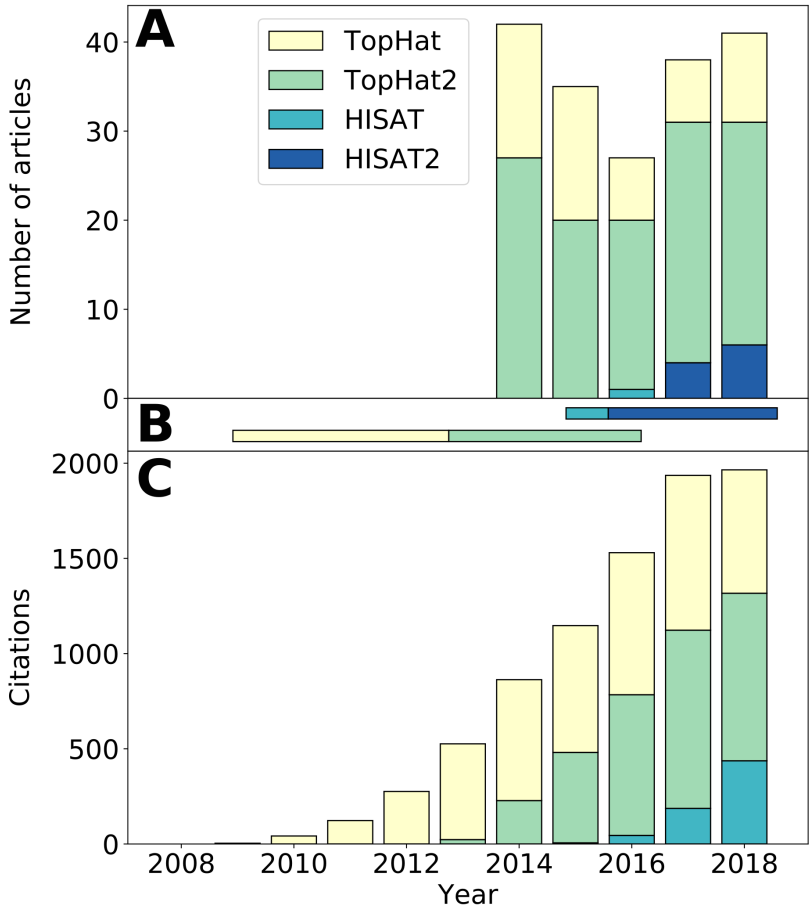


Figure 3 – Observed latency in tool usage — a TopHat–HISAT case study. **A** illustrates the distribution of articles using tools from the TopHat–HISAT family found in our methodological literature review. **B** presents the recommended usage period for each tool. Dates were extracted from the TopHat and HISAT pages, using official release dates and notices given by the authors. **C** represents the distribution of new citations per year for each software original publication. The citation count was extracted from Scopus in January 2019. HISAT and HISAT2 share the same color considering that HISAT2 was never published independently of HISAT. While **A** only includes articles using RNA-seq with Homo sapiens, **C** includes all articles citing one of the tools.

2.3.3 RNA-seq methodology is incompletely reported with some steps much better described than others

Because several different references and software tools are used for each of the different RNA-seq analysis steps, the number of resulting pipelines is combinatorially large, increasing the importance of ensuring the complete description of all steps. Unfortunately, among the 465 articles considered, many failed to even mention any information regarding an RNA-seq analysis step (no information shares in Figure 2). An interesting observation is the non-uniform distribution of this lacking information. If the distribution could be explained by a random phenomenon or the presence of a bimodal distribution of articles ranked by a

global ‘reporting quality’ metric, one would observe approximately the same share missing from each of the steps described in Figure 2, which is not the case. This non-uniform distribution could be interpreted as mirroring the community-conceived importance of the different steps. For example, the alignment tool is specified in 88% of the articles, whereas the preprocessing tool is indicated in only 24%. This could mean that the community of RNA-seq users believes that the alignment is far more important than the preprocessing in an RNA-seq experiment. The vast differences highlighted for preprocessing could also be explained by RNA-seq workflows that do not include a preprocessing step, or by prior preprocessing by the sequencing facility. In either case, the manuscript should contain all relevant information about preprocessing. The field currently lacks a meaningful quantitative assessment comparing the inherent bias of the different steps to the biological background. With such information, it would be possible to rank the importance of each decision on the final quantification, and appreciate if those are in relationship with the methodological results shown here. In any case, display of information for all analysis steps of an RNA-seq experiment is an important goal that could be achieved through awareness and enforced publication guidelines.

We also note that even when a step is described in an article, information is often missing regarding versions or parameter values, which are necessary to ensure reproducibility. Such is the case, for example, for the step specifying the genomic annotation, for which one can observe a non-uniform distribution in the proportion of articles giving details about the version number (Figure 2D, differences in the inner donut chart shading). In particular, most articles using Ensembl [15] and GENCODE [16] annotations specify the version number while few articles using RefSeq [17] do. A possible explanation for this is the availability of such information at the moment users download the genome annotation. A clear display of the annotation version for the different species by RefSeq on their website could help it reach the same level of version specification as Ensembl and GENCODE. In general, an upfront display of version information is crucial for any database, which could be subject to modifications. Any information that is easily accessible seems to bear more importance than information that is harder to find. One also has to be able to access previous versions of a reference to keep older datasets relevant. While Git is not currently designed to support large data files, Git-like features would answer these goals, by providing the possibility to navigate across the different versions of a reference, explicitly versioning data files and their differences. Current scientific data hosting platforms (e.g. figshare, Git LFS, Open Science Framework, Quilt, Zenodo) do not support either versioning or diffing of files, which is necessary for efficient data tracking.

Another possible cause for the unreliable distribution of information is the availability of alternative sources for the same references. Genome sequences and annotations can be readily downloaded from their official maintainer's website, but can also be found on other file transfer protocol (FTP) servers or websites. Duplication of information only increases the risk of out-of-date data and versioning errors.

To summarize the findings, Figure 4 presents how many essential steps were correctly specified per article. This provides an idea of how the lack of information is distributed in the literature. While Figure 4A, showing the distribution of articles according to the number of steps they described, offers an optimistic view of article methodological quality, Figure 4B, displaying the distribution of articles classified by the number of steps they completely describe, represents the actual reproducibility potential. It is worth noting that some articles display absolutely no information other than the fact they have performed RNA-seq. Supplementary Figure 1 illustrates the completeness of each essential criterion per year. We do not note any obvious improvement of the situation in recent years.

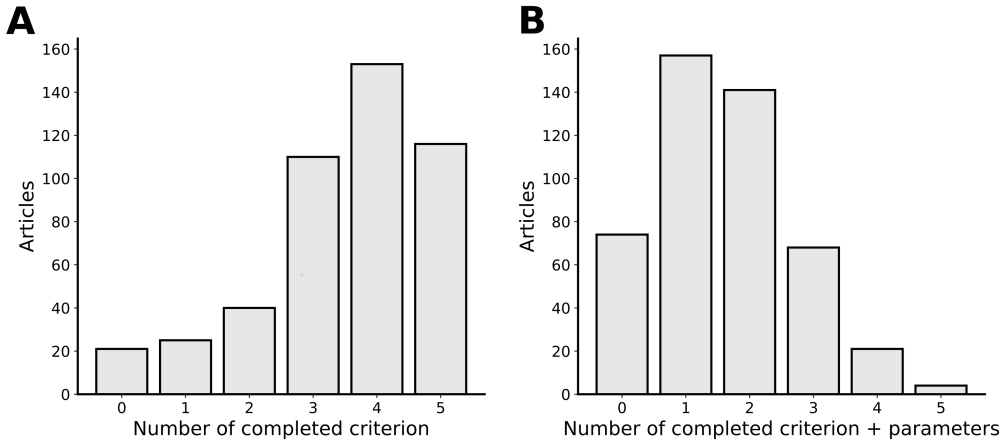


Figure 4 – Article distribution by completeness. **A.** Distribution of articles by the number of essential criteria that have been specified in the methodology. Essential criteria are considered to be the dataset, alignment type, genomic annotation, alignment tool and quantification tool. **B.** A criterion needed to have every parameter specified to be accepted.

2.3.4 Measures that could be undertaken to increase methodological reporting of RNA-seq computational pipelines

Guidelines for adequate information presentation in RNA-seq already exist. The MINSEQE, minimum information about a high-throughput sequencing experiment, analogous

to the former minimum information about a microarray experiment [18], describes the required information 'to enable the unambiguous interpretation and facilitate reproduction of the results of the experiment'. Results of the methodological review suggest that such guidelines are routinely not being followed. A more reliable way to enforce a desired behavior would be at the publishing level, with editors requiring standardized minimal information about an RNA-seq experiment. Nevertheless, it would not be enough. The problem is also due to the way information is presented. Textually describing the experiments is a potential source of loss of clarity. The bias-less way to report an RNA-seq in silico experiment is probably a direct access to the code used to generate the results. If this code is housed in a collaborative version control system (e.g. Git), one would also be able to update the code for the peer-reviewing step or additional corrections. In fact, Git is already used for software version control and even scientific paper writing and peer-reviewing [19]. But this solution suffers from potential code readability and computing infrastructure issues. In lieu of an in silico physical preservation of a workflow, a semantical one offers workflow reproducibility, with less infrastructure dependency [20]. A schematic and semantic view of the data transformation pipeline, as already proposed in other areas of scientific computing [21], would help to better illustrate every data and software linkage. In this view, we advocate the use of workflow management systems, such as Nextflow [22] and Snakemake [23]. These tools promote reproducibility by explicitly defining and compartmentalizing the different pipeline steps, and enabling a scalable execution of the pipeline in dedicated containers and virtual environments. Users can therefore publish more readily readable and reproducible code, all while following previously described rules for reproducible computational research [24].

2.4 Conclusion

In summary, we illustrate the lack of information, the unreliable distribution of references and the latency in software usage in the RNA-seq literature by the means of a methodological review of the literature. The current state of the literature prohibits meaningful meta-analysis of the literature and large-scale reproducibility studies. We believe this situation will be improved by acknowledging the issue, clearly displaying the technical requirements for RNA-seq methodological reproducibility and with scientific publishers demanding standardized, high-quality methodology [25].

2.5 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-05412 to M.S.S.). J.S. was supported by a Masters scholarship from NSERC. M.S.S. holds a Fonds de Recherche du Québec—Santé Research Scholar Junior 2 Career Award. M.S.S. is a member of the Centre de Recherche du Centre Hospitalier de l'Université de Sherbrooke.

2.6 References

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray.," *Science (New York, N.Y.)*, vol. 270, pp. 467–70, oct 1995.

[2] S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, "Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells," *PLoS ONE*, vol. 9, p. e78644, jan 2014.

[3] J. Ison, K. Rapacki, H. Ménager, M. Kalaš, E. Rydza, P. Chmura, C. Anthon, N. Beard, K. Berka, D. Bolser, T. Booth, A. Bretaudeau, J. Brezovsky, R. Casadio, G. Cesareni, F. Coppens, M. Cornell, G. Cuccuru, K. Davidsen, G. D. Vedova, T. Dogan, O. Doppelt-Azeroual, L. Emery, E. Gasteiger, T. Gatter, T. Goldberg, M. Grosjean, B. Grüning, M. Helmer-Citterich, H. Ienasescu, V. Ioannidis, M. C. Jespersen, R. Jimenez, N. Juty, P. Juvan, M. Koch, C. Laibe, J.-W. Li, L. Licata, F. Mareuil, I. Mičetić, R. M. Friborg, S. Moretti, C. Morris, S. Möller, A. Nenadic, H. Peterson, G. Profiti, P. Rice, P. Romano, P. Roncaglia, R. Saidi, A. Schafferhans, V. Schwämmle, C. Smith, M. M. Sperotto, H. Stockinger, R. S. Vařeková, S. C. Tosatto, V. de la Torre, P. Uva, A. Via, G. Yachdav, F. Zambelli, G. Vriend, B. Rost, H. Parkinson, P. Løngreen, and S. Brunak, "Tools and data services registry : a community effort to document bioinformatics resources," *Nucleic Acids Research*, vol. 44, pp. D38–D47, jan 2016.

[4] K. D. Hansen, Z. Wu, R. A. Irizarry, and J. T. Leek, "Sequencing technology does not eliminate biological variability.," *Nature biotechnology*, vol. 29, pp. 572–3, jul 2011.

[5] J. L. Spudich and D. E. Koshland, "Non-genetic individuality : Chance in the single cell," aug 1976.

[6] P. L. Auer and R. W. Doerge, "Statistical design and analysis of RNA sequencing data.," *Genetics*, vol. 185, pp. 405–16, jun 2010.

[7] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq : an assessment of technical reproducibility and comparison with gene expression arrays.," *Genome research*, vol. 18, pp. 1509–17, sep 2008.

[8] M. A. Busby, C. Stewart, C. A. Miller, K. R. Grzeda, and G. T. Marth, "Scotty : a web tool for designing RNA-Seq experiments to measure differential gene expression," *Bioinformatics*, vol. 29, pp. 656–657, mar 2013.

[9] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton, "How many

biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use ?," *RNA (New York, N.Y.)*, vol. 22, pp. 839–51, jun 2016.

[10] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, no. 1, pp. 1–19, 2016.

[11] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq : a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, jan 2009.

[12] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat : discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, pp. 1105–1111, may 2009.

[13] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2 : accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, p. R36, apr 2013.

[14] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT : a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, pp. 357–360, apr 2015.

[15] D. R. Zerbino, P. Achuthan, W. Akanni, M. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek, "Ensembl 2018," *Nucleic Acids Research*, vol. 46, pp. D754–D761, jan 2018.

[16] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang,

B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek, "GENCODE reference annotation for the human and mouse genomes," *Nucleic Acids Research*, vol. 47, pp. D766–D773, jan 2019.

[17] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (RefSeq) database at NCBI : current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research*, vol. 44, pp. D733–D745, jan 2016.

[18] A. Brazma, "Minimum Information About a Microarray Experiment (MIAME)–successes, failures, challenges.," *TheScientificWorldJournal*, vol. 9, pp. 420–3, may 2009.

[19] D. S. Katz, K. E. Niemeyer, and A. M. Smith, "Publish your software : Introducing the Journal of Open Source Software (JOSS)," *Computing in Science & Engineering*, vol. 20, pp. 84–88, may 2018.

[20] I. Santana-Perez, R. Ferreira da Silva, M. Rynge, E. Deelman, M. S. Pérez-Hernández, and O. Corcho, "Reproducibility of execution environments in computational science using Semantics and Clouds," *Future Generation Computer Systems*, vol. 67, pp. 354–367, feb 2017.

[21] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, "Examining the Challenges of Scientific Workflows," *Computer*, vol. 40, pp. 24–32, dec 2007.

[22] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature Biotechnology*, vol. 35, pp. 316–319, apr 2017.

[23] J. Koster and S. Rahmann, "Snakemake–a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, pp. 2520–2522, oct 2012.

[24] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten Simple Rules for Reproducible Computational Research," *PLoS Computational Biology*, vol. 9, p. e1003285, oct 2013.

[25] J. Simoneau and M. S. Scott, "In silico analysis of RNA-seq requires a more complete description of methodology," *Nature Reviews Molecular Cell Biology*, p. 1, may 2019.

2.7 Suplementary data

2.7.1 Supplementary methods

Methodological review. The methodological review procedure used was similar to a systematic review, but without a full review of the relevant literature, due to the large quantity of articles using RNA-sequencing. First, a subset of articles was selected from the literature. An arbitrary number of 1000 articles from the last 5 years was selected to perform the study. These articles were chosen using Scopus and the following query : "(TITLE-ABS-KEY(align* OR mapp* OR annotation OR genom* OR transcript* OR reference OR dataset) AND TITLE-ABS-KEY("RNA-seq*" OR "RNA seq*") AND TITLE-ABS-KEY(biops* OR patient* OR cell*) AND TITLE-ABS-KEY(human*))". The results were arranged by the year of publication, and the 200 newest articles were selected for each year. After the article selection, two reviewers independently assessed whether the articles respected the inclusion criteria and proceeded with the data extraction if applicable. At the end of the independent review, the reviewers compared their results. Identical results were kept as is. Discrepancies were further investigated by the two reviewers until a consensus was reached.

Article inclusion criteria. The detail about articles inclusion or rejection is displayed in Supplementary Table 1. The several rejection tags are described here. The main motivation for the design choices regarding the article inclusion was to only keep articles with comparable pipelines. The study was limited to articles using RNA-seq (R1) as an experimental tool to study Homo sapiens (R2). To be included, the article needed to encompass the in silico pipeline from FASTQ files to quantification, meaning that articles using processed read counts (R3) or lacking quantification (R5) were rejected. Articles that were studying specific steps of the RNA-seq in silico pipeline were also rejected due to alteration of the pipeline (R4). The study was also limited to short-read sequencing due to the different software requirement of other technologies (R6). Finally, articles that were note accessible through our university access resources (R7), or that are retracted (R8) were rejected. Of the 1000 articles originally selected, 465 were kept based of these criteria.

Data extraction. Reviewers curated specific information about the RNA-seq pipeline, following steps and parameters of interest defined in Fig. 1. Only applicable and truthful information was extracted, e.g. if an inexistent software version number was specified in the article, it was ignored. Contradictory and confused information were also deemed unreliable and were thus ignored. Complete results are available as Supplementary data.

2.7.2 Supplementary figures and tables

Tags	Article status	N
A	Accepted	465
R2	Rejected : Non-human RNA-seq experiment	209
R1	Rejected : No RNA-seq experiment	209
R3	Rejected : Processed RNA-seq data	67
R4	Rejected : RNA-seq meta-analysis	29
R5	Rejected : No quantification	15
R6	Rejected : PacBio long reads	2
R7	Rejected : No access	2
R8	Rejected : Retracted	2

Table 2.1 – Distribution of the articles. Classification of the articles used for the review. Only information contained in the accepted articles was considered in the study.

Chapitre 3
Article 2 : Factorial study of the RNA-seq computational workflow identifies biases as technical gene signatures

Factorial study of the RNA-seq computational workflow identifies biases as technical gene signatures

Auteurs de l'article: Joël Simoneau, Ryan Gosselin, Michelle S. Scott

Statut de l'article: Article soumis dans *NAR Genomics and Bioinformatics* ;
Article publié dans bioRxiv. 2020 Jan 30 ; [DOI: 10.1101/2020.01.30.924092]

Avant-propos: Partant des informations d'usage méthodologique, le deuxième article caractérise les biais de quantification principaux retrouvés dans la littérature. Des ensembles de données sont traités avec diverses combinaisons de pipelines constitués des outils et références les plus utilisés dans la littérature, et des ensembles de gènes permettant la classification des résultats en fonction de la méthodologie utilisée sont identifiés. La comparaison des biais en fonction des étapes expérimentale démontre que l'annotation génomique et l'outil de quantification sont les deux choix ayant le plus d'impact sur les résultats.

Contributions : JS a effectué l'analyse des données. JS, RG et MSS ont conceptualisé la recherche et participé à la rédaction du manuscrit.

Résumé

Le séquençage de l'ARN est une technologie modulaire combinant des approches expérimentales et computationnelles, dans le but d'identifier et de quantifier des molécules d'ARN. La modularité du pipeline de séquençage de l'ARN permet l'adaptation des protocoles afin de développer de nouvelles approches pour explorer la biologie de l'ARN. Par contre, cette modularité crée le besoin d'une méthodologie rigoureusement et exhaustivement rapportée. De la liberté méthodologique vient la responsabilité de faire des choix éclairés. Ici, nous présentons une approche permettant d'identifier des biais de quantification affectant spécifiquement des groupes de gènes partageant des caractéristiques communes dans le contexte actuel des outils logiciels et des références utilisés en séquençage de l'ARN. Pour ce faire, nous avons généré et utilisé une large gamme combinatoire de pipelines computationnels, puis décomposé les données d'expression en utilisant une méthode de factorisation de matrice nommée la décomposition en composantes indépendantes. En étudiant le pipeline de séquençage de l'ARN en utilisant cette approche systématique, nous mettons en lumière l'importance souvent non caractérisée des annotations génomiques dans les résultats de quantification. Nous démontrons aussi que les différents choix méthodologiques du pipeline ne sont pas indépendants, illustrés par une interaction entre les annotations génomiques et les logiciels de quantification. Les biais de quantification des gènes semblent principalement venir des différences de séquences, de la présence de gènes partageant des coordonnées chromosomiques et par des gènes ayant des séquences similaires. Notre approche offre une explication aux biais observés en identifiant les caractéristiques communes des gènes affectés, caractéristiques qui sont différentiellement utilisées, ou considérées, par les différents logiciels. Cette information offre des avenues pour la création d'outils plus informés.

3.1 Abstract

RNA-seq is a modular experimental and computational approach that aims in identifying and quantifying RNA molecules. The modularity of the RNA-seq technology enables adaptation of the protocol to develop new ways to explore RNA biology, but this modularity also brings forth the importance of methodological thoroughness. Liberty of approach comes with the responsibility of choices, and such choices must be informed. Here, we present an approach that identifies gene group specific quantification biases in currently used RNA-seq software and references by processing sequenced datasets using a wide variety of RNA-seq computational pipelined, and by decomposing these expression datasets using an independent component analysis matrix factorisation method. By exploring the RNA-seq pipeline using a systemic approach, we highlight the yet inadequately characterized central importance of genome annotations in quantification results. We also show that the different choices in RNA-seq methodology are not independent, through interactions between genome annotations and quantification software. Genes were mainly found to be affected by differences in their sequence, by overlapping genes and genes with similar sequence. Our approach offers an explanation for the observed biases by identifying the common features used differently by the software and references, therefore providing leads for the betterment of RNA-seq methodology.

3.2 Introduction

Modularity is both a boon and a burden for RNA-sequencing (RNA-seq) analysis. At its core, RNA-seq leads to the identification and quantification of RNA molecules from a biological extract [1]. RNA-seq is in fact an umbrella term, encompassing a broad diversity of laboratory and computational design choices, where each choice defines the scope of the study, the questions it might answer [2]. The modularity of RNA-seq has been a steppingstone in the development of many other techniques, mainly differing by the way in which the RNA is extracted, and with consequent modifications of the in silico pipeline. For example, ribosome profiling (Ribo-seq) can be summarized as the RNA-seq of the RNA fragments protected by ribosome footprints [3]. The modularity of the RNA-seq in silico pipeline, stemming from the usage of well-defined data processing steps, each supported by specific file formats, has led to the creation of specialized software for each of the different steps, compartmentalizing the data processing. Having defined data processing steps helps to isolate each technical problem, creating an ecosystem where research groups specialize

in answering individual steps, and proceed to benchmark accordingly.

Due to the RNA-seq modular nature, one has to constrain many degrees of freedom linked to the experimental design to be able to generate and process the data. These degrees of freedom represent protocols, reagents and kits on the experimental side, and software, references and parameters on the computational side. With a benchmarking point of view, experimental and computational design choices mainly differ by their permanence. While a specific RNA extract can only be processed once in the laboratory, sequencing being a destructive method, sequencing data can theoretically be reanalyzed in a wide variety of ways. This facilitates the creation of large-scale benchmarking of the RNA-seq in silico pipeline, because they can be built on a very specific and unique group of datasets, while benchmarking studies of the RNA-seq laboratory pipeline have to deal with extra noise and reproducibility issues coming from biological variability. RNA-seq in silico pipeline benchmarking studies are also cheaper to produce since they only require computational resources.

Benchmarking the RNA-seq pipeline may take two different approaches, either analytical or systemic. In an analytical approach, the pipeline would be studied in its irreducible form, meaning that each step would be benchmarked independently. The analytical approach is usually used when publishing a new tool. To ensure the relevance of the newly proposed method, authors will compare its performance to current methods [4]. Depending on the position of the studied step in the overall pipeline, it might be difficult to meaningfully assess its quality. For example, alignment software has often been characterized using the percentage of alignment as a metric to be optimized, even though such a metric does not hold any biological meaning [5]. To bypass the need for other metrics, it is also possible to study the effects of a given step on the rest of the pipeline, using a fixed downstream processing. Conversely, in a systemic approach, the pipeline is studied as a whole. While the analytical approach is based on the hypothesis that each step is fully independent, the systemic approach can be used to study interactions between steps.

		Fonseca, 2014	Robert, 2015	Germain, 2016	Everaert, 2017	Sahraeian, 2017	Williams, 2017	Zhang, 2017	Costa-Silva, 2017	Baccarella, 2018	Merino, 2019	This study
Trimmer												
Cutadapt	[24]											●
Trimmomatic	[25]					●						●
Annotation												
Ensembl	[32]	●	●		●	●						●
GENCODE	[79]					●		●	●			
RefSeq	[33]			●								●
Aligner												
Bowtie	[56]	●									●	
Bowtie2	[57]	●		●	●		●		●			
BWA	[58]	●							●			
BWA-SW	[59]	●										
GSNAP	[60]	●										
HISAT	[61]			●								
HISAT2	[27]					●	●			●		●
MapSplice	[62]			●								
OSA	[63]	●										
SeqMap	[64]						●					
SMALT	-	●										
STAR	[28]	●	●	●		●	●	●	●	●		●
TopHat	[65]	●										
TopHat2	[26]	●	●	●	●	●	●		●		●	●
Quantifier												
BEDTools	[66]										●	
BitSeq	[67]						●					
Cufflinks	[49]	●				●		●				
Cufflinks2	[29]	●	●	●	●		●				●	●
DEXSeq	[68]										●	
eXpress	[69]					●	●	●				
featureCounts	[30]			●		●						●
FluxCapacitor	[70]	●										
HTSeq	[31]	●	●		●		●		●			●
IsoEM	[71]						●					
RSEM	[72]			●			●	●		●	●	
rSeq	[73]						●					
STAR	[28]								●	●		
StringTie	[74]					●	●			●		
TIGAR2	[75]							●				
Pseudo-aligner												
Kallisto	[76]			●	●	●	●	●	●	●		
Sailfish	[77]		●	●		●	●	●				
Salmon	[78]			●	●	●	●	●	●	●		

Table 1 – Compilation of software and genome annotations used in articles benchmarking at least two different steps of the RNA-seq in silico pipeline. Software is classified by steps, where pseudo-aligners are considered separately because they overlap more than one step. Major re-release of software with an independent publication is considered as a separate software. The last column describe softwares considered in the present study.

In Table 1, we compiled a list of articles benchmarking the RNA-seq in silico pipeline by considering more than one step in their analysis [6–15]. The main point of interest is the imbalance between the pipeline steps, both in the number of elements studied per article as well as globally. These articles often use the term "RNA-seq workflow" to describe the object of their analysis, while also mainly limiting themselves to the alignment and quantification software. As we described earlier, the RNA-seq in silico workflow should consider all steps from the raw FASTQ files to the count matrices [16]. Insofar as we do not have any study highlighting the importance, or lack of, of every workflow step, overlooking some steps might hide important biases. In a previous study, we highlighted that the trimming step and the choice of genomic annotations were often not reported in methodologies of articles performing RNA-seq [17]. Furthermore, we can observe that these steps are also overlooked in the articles studying the RNA-seq workflow (Table 1). Only one article included in our analysis reported using more than one annotation reference, and only used it to evaluate transcript assembly. Many of those benchmarks also did not provide any information about trimming. We now find ourselves before a circular causality problem in which we are not benchmarking certain steps because they are not being reported and we are not reporting them because they are not being benchmarked. In any case, there is insufficient data for a meaningful answer regarding whether the overlooked information holds any importance.

Considering the difficulty in obtaining a high-quality gold standard for all genes in an RNA-seq study, we propose another strategy to identify biases in the processing pipeline. Instead of assessing the divergence in relation to the ground truth, we suggest treating this as a classification problem. If we process datasets with a variety of different pipelines, and then find some gene signatures classifying the processed datasets in accordance to some pipeline choice, then we would have identified processing biases affecting the quantification of genes.

Matrix factorisation methods are important tools for data-driven analysis, used to identify the main characteristics of highly dimensional datasets. Principal Component Analysis (PCA) is usually the go-to method used when confronted with such problems. PCA deterministically explains the variance of a dataset by projecting it onto decreasingly important latent variables, each constrained to be orthogonal to one another, uncorrelated. However, we chose to apply another method in this article, namely Independent Component Analysis (ICA). ICA decomposes a dataset into a specified number of latent variables of unconstrained variance, while optimizing for their independence, i.e. minimizing their mutual

information. Both techniques produce uncorrelated latent variables, but only ICA produces independent latent variables when the original signal is non-Gaussian [18]. Where PCA highlights the largest trends present within the dataset, ICA seeks to extract independent structures, or phenomena, occurring within the dataset. This is due to a significant difference in the hypotheses at the core of these methods. PCA considers the data to follow a multivariate Gaussian distribution whereas ICA seeks a linear combination of non-Gaussian distributions. ICA has previously been applied to RNA-seq quantification results to infer groups of genes displaying a shared behaviour across several datasets [19, 20]. ICA is a long-sought answer to the cocktail party problem, where an unknown number of persons talk in a room where a known number of microphones are placed [21]. The goal of the problem is to decorrelate the different microphone feeds to isolate the original speech of each person. The main hypothesis of an ICA is that every observation, microphone input, is a linear combination of a set of sources, herein the persons. In our case, the expression of genes acts as our microphone feeds, and each person can be thought of as a cellular process, a level of regulation, or a technical bias which has effect on the expression level of a subset of genes. By decomposing RNA-seq datasets generated using different treatments or biological origins, one can identify the sources, named expression modes in the context of RNA-seq studies [19], that are important in defining and differentiating our datasets.

The biological importance of expression modes can be inferred by correlating them with known biological features. ICA has also been previously used to identify and remove batch effects by correlating expression modes with experimental features within the datasets [22]. We will dub the expression modes as either biological or technical modes, based on the types of variables with which they correlate. While technical modes correlating with sequencing batch might not hold information of interest for us, we hypothesize that this could be used as a tool to study biases in the RNA-seq in silico processing pipeline. By analyzing a number of datasets using a wide variety of different software and references, we could identify technical modes classifying the datasets by the pipeline used. These hypothetical technical modes would not appear in a normal RNA-seq experiment, where we do not usually use multiple software in parallel to accomplish the same step.

In this study, we processed biological replicates of different human tissues with a wide range of RNA-seq in silico pipelines obtained from the exhaustive combinations of selected software and genome annotations. We primarily chose software and references that are currently reported as being used in the RNA-seq literature, in order to represent the present situation [17]. We then decomposed these expression data into expression modes using an

ICA analysis. We further characterized these modes as either biological or technical modes, on the basis of the variables that they can classify. Technical modes were then studied to explain the observed bias, identifying the features responsible for a different behavior of the software. A differential gene expression analysis was also produced for the different pipeline steps, highlighting the number of genes globally affected by these steps.

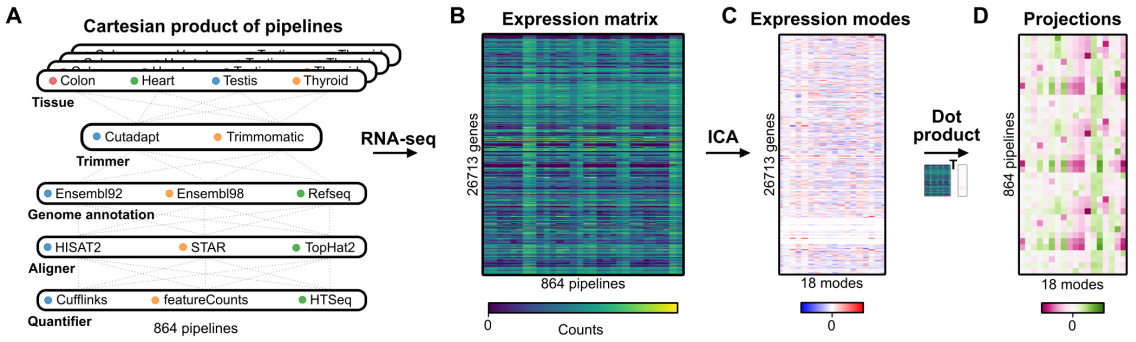


Figure 1 – Illustrations of the main steps of this study. First, all possible combinations of tissue samples, trimmers, genome annotations, aligners and quantifiers are processed as independent RNA-seq experiments. These results are compiled in an expression matrix which is decomposed into expression modes using an ICA. Projections, used to identify the information contained in the expression modes, is generated as the dot product of the expression matrix and the expression mode matrix.

3.3 Methods

3.3.1 Cartesian product of RNA-seq workflows

We used previously published RNA-seq datasets of human tissues from Array-Express E-MTAB-2836 [23] (https ://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2836/). Samples used are available in Supplementary Table 1. In order to have different levels of in-group and between-group variability, we chose four different tissues (colon, heart, testis, thyroid), each represented by four samples coming from different individuals. To evaluate the impact of every pipeline step, we processed all the datasets with a wide variety of RNA-seq workflows. We chose either recent, or commonly used, software and genomic annotations for each of the different RNA-seq steps considered, as defined in Figure 1. To keep every step independent from one another, we chose to exclude any software that encompasses more than one step (i.e. pseudo-aligners that overlap the alignment and quantification steps). We processed each dataset with the full compendium of Cartesian products of pipeline choices, meaning every possible unique combination of software

and references. Using a design of experiments (DOE) terminology, this represents a full factorial experiment. To keep a basis of comparison, we used gene level counts as the output of the different pipelines. We only used tools which directly report gene counts, so there is no transformation of the final output.

3.3.2 RNA-seq methodology

We performed RNA-seq using only methods that rely on genome-based alignment, and software that are restricted to a single methodological step. We kept default parameters for most of the options, as to mimic what is being done in the literature [17]. FASTQ files were downloaded from the SRA repository and trimmed independently using Cutadapt [24] v2.3 and Trimmomatic [25] v0.36. For the trimming parameters, we used a minimal Phred quality score of 15, and kept only reads that were at least 75 nucleotides after trimming, knowing that we are working with unstranded paired-end reads of 100 nucleotides. The alignment was performed independently using TopHat2 [26] v2.1.1 (wrapping Bowtie2 v.2.3.4.3), HISAT2 [27] v2.1.0 and STAR [28] v2.5.3a. The aligners were run with default settings for unstranded data. It is important to note that only STAR requires an annotation file at this point. The other two software were not provided an annotation file for the alignment. The quantification was performed independently using Cufflinks [29] v2.2.1, featureCounts [30] (Subread v1.6.4) and HTSeq [31] v.0.11.2. Quantification was summarized as gene-level counts. Ensembl [32] version 92 and 98, and RefSeq [33] release 109 were used as the different genome annotations. Ensembl 92 and RefSeq 109 were both released in April 2018, making them temporally comparable. Ensembl 92 and Refseq 109 are both built upon the GRCh38.p12 genome, while Ensembl 98 uses GRCh38.p13 genome. Because the primary assembly of both these reference genome versions is the same, and because we restricted our studied genes to the primary assembly, only GRCh38.p13 was used. Every time that a genome annotation was needed in a step, this step was processed three times, one with each annotation. The detection of differentially expressed genes (DEGs) was performed with DESeq2 [34] v1.26. All software tools were installed locally through Bioconda [35]. The dependencies and parameters for each pipeline steps are accessible in a Snakemake [36] project.

3.3.3 Data preprocessing

Raw counts from the different pipelines were combined in one expression matrix. Due to the fact that we are using more than one genome annotation, we require a common identifier to compare genes from Ensembl and RefSeq. To do so, we used the HUGO Gene Nomenclature Committee (HGNC) resource [37]. We considered data in the HGNC resource that were provided by Ensembl and the NCBI, while prioritizing information for HGNC in case of conflict. This also means that all results presented in this work are only drawn upon genes that are present in HGNC, ignoring genes that are unique to a specific genome annotation. After filtering for genes present in HGNC and quantified through the several pipelines, we are left with 26 713 genes. Raw counts were also preprocessed before being fed into the ICA model. For the first preprocessing step, we used the varianceStabilizingTransformation (fitType='local') function from the DESeq2 project [38]. This step scales the different experiments so that they all have the same weight, and ensures the homoskedasticity of the genes, meaning that the variance of the genes is not function of their expression level. Homoskedasticity is important because the biological importance of a gene is not directly linked to its absolute expression value, and without this correction, the dataset features would be largely driven by a small number of highly expressed genes. The expression matrix is then transformed by a Mahalanobis whitening, rotating the dataset to decorrelate the different dimensions [39].

3.3.4 ICA model

We used the scikit-learn implementation of FastICA to process our dataset [40]. FastICA maximizes the neg-entropy, a measure of the non-Gaussianity of the components [21]. This optimization is performed using an iterative method, requiring the user to specify a tolerance, i.e. the minimum change of neg-entropy needed to stop the iterations. Because we need to perform FastICA with different numbers of components and due to the fact that the neg-entropy measure scales with the number of components, choosing a sensible tolerance is not trivial. A tolerance that is too large would end the optimization early, without attaining the real maximum, while a tolerance too small would never end the optimization. In order to avoid obtaining spurious maxima, we choose to force the FastICA algorithm to work with a really small tolerance (iteration step tolerance of 1e-18), and a large number of maximum iterations (1e5 iterations). While preventing the algorithm from stabilizing, we ensure that the optimization does not stop prematurely.

3.3.5 ICA stability and independence

We then confirmed the robustness of the optimization maximum and the independence of the components. To do so, we ran the FastICA multiple times (n=25), using different starting points for the optimization. Afterwards, we computed the correlation matrix for the different components. In theory, an optimal correlation matrix for this problem should be a block identity matrix, where each block is a square with the same length as the number of replicates. Figure 2A contains an example of a near optimal matrix (with M=18), and two inadequate matrices (M11 and M26). The uniformity of the blocks confirms that the FastICA has found the same maximum for the different runs, and the identity matrix, where off-diagonal elements are zero, confirms the independence of the components. We scored the correlation matrix by quantifying its divergence from the optimal correlation matrix using the mean squared error (MSE).

3.3.6 Identifying expression modes

To identify what information an expression mode is providing, we used a k-nearest neighbour (KNN) classification approach. We first needed to generate projections of the pipelines along the expression modes. As in Figure 1, the projections are calculated as the dot product of the expression matrix by the expression modes matrix. This provided us with a one-dimensional distribution of the pipelines along each expression mode. We then quantified whether the pipelines are clustered according to biological or technical variables. To do so, using one projection at a time, and for each pipeline, we quantify the percentage of the 50 nearest pipelines, in terms of distance along the expression mode projection, that share the same label as the pipeline of interest for a specific biological or technical variable. This was done for all the different methodological choices from the different pipeline variables, taking the average percentage for each choice. This score informs us about the uniformity of the clusters found in a projection.

Each expression mode is defined by attributing a weight to each gene (Figure 1C), where genes with extreme values contribute more to the definition of the mode. In order to work with a list of genes, we needed to find a weight threshold at which genes would be considered as a part of the expression mode. We selected genes that were farther than four standard deviations from the distribution average, which creates gene groups with approximately 30 to 300 genes. The selected genes and their weights for all expression

modes are available in Supplementary Data 1 for the original model and Supplementary Data 2 for the Cufflinks-only model. Only weights outside the four standard deviations were kept, the remainder were transformed to a zero value.

3.4 Results

3.4.1 ICA highlights biological and technical differences between the RNA-seq pipelines

To run an ICA decomposition, one has to specify a number of expression modes (M) to be generated. This number is an unknown parameter and varies according to the underlying structure of the dataset. In order to identify the optimal number of expression modes to represent our dataset, we performed the ICA with a wide range for M (6 to 35), and we quantified the stability and the independence of the expression modes for these models. Figure 2A illustrates the distribution of the mean squared error (MSE) over the different number of expression modes used. Several values of M seem to be suitable for analysis, with a MSE of approximately zero. We chose to analyze the model with M = 18, being the model with the largest number of expression modes, while having the smallest MSE found. We favoured the largest number of stable expression modes with the hypothesis that decomposing the same dataset into more components would mean that the resulting components would be simpler, less convoluted.

We then identified the information given by each expression mode using the KNN score, displayed in Figure 2B. The heatmap is separated in five different blocks, each representing a pipeline variable, with the different choices, included in this study. For each block, the minimum possible score is 100% divided by the number of elements in that block, which is a score equivalent to random guessing. The heatmap should be read in a column-wise manner, looking at what information each expression mode is providing. The majority of the expression modes seem to be driven by biological information, which are the modes that are usually studied when using ICA with RNA-seq data [20, 41] and the modes of interest for researchers using RNA-seq to gain insight into biology. To confirm the informational value of the biological modes, we illustrated the distribution of the pipelines along four selected biological modes, one for each tissue, in Figure 2C. We can observe that each distribution offers a clear separation for a specific tissue, meaning that the underlying gene weights can be used as a biological gene signature for tissue classification. The pairwise comparisons also confirm that the modes are independent from one another because the

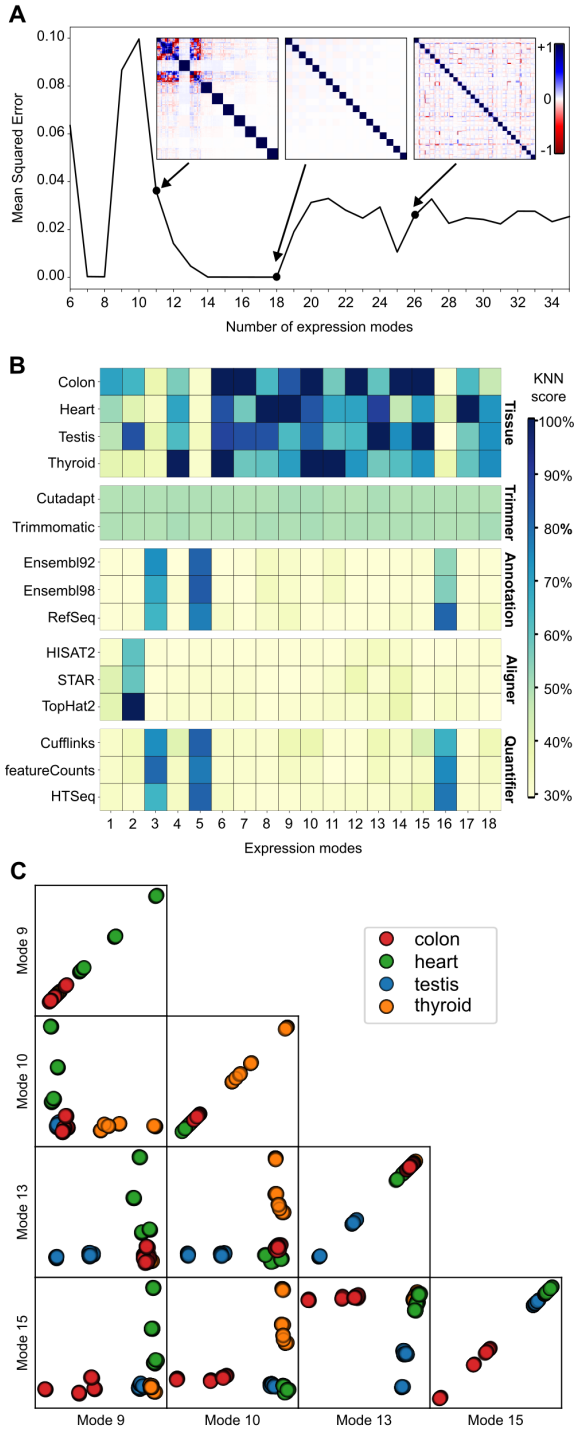


Figure 2 – General information about the processed ICA model. **A** represents the Mean Square Error (MSE) for ICA model computed with different numbers of expression modes (M). The MSE is calculated using the theoretical optimal block diagonal matrix. Covariation matrix for models with M of 11, 18 and 26 are shown. **B** illustrates the information given by each expression mode. The heatmap is separated into five different blocks, each representing one variable choice. A high KNN score means that the variable is well clustered and well separated from the other variables. **C** illustrates distribution and pairwise distribution for four biological modes, one for each tissue type.

pipelines are distributed as two perpendicular lines. We further characterized these gene groups using a Gene Ontology enrichment analysis provided in Supplementary Figure 1. These figures show that the genes used to classify the different tissues are also biologically related to the tissues, meaning that we have learned from biologically relevant features of the dataset. While the first block of Figure 2B represents the general use of ICA, i.e. studying biological features of a dataset, and serves as a positive control for our approach, our interest lies in the four other blocks.

The choice of trimming software, studied here using Cutadapt and Trimmomatic with the same set of parameters, does not have any impact on the dataset that could be identified by the ICA. The trimming block in Figure 2B shows a uniform block of value 50% for every expression mode, meaning that the distribution of the pipeline along the expression modes is purely random. While this does not prove that the trimming software does not have any impact, it shows that this impact would be smaller than, and therefore hidden by, the other expression modes.

The choice of alignment software is captured in the expression mode 2 (EM2), where TopHat2 is shown to be fully separated from the other software, whereas HISAT2 and STAR seem to be indissociable from one another, due to their similar scores. Genome annotations and quantification software are interlinked in three different technical modes (EM3, EM5 and EM16). Detecting technical modes is only the first part of the problem. Having shown that an ICA decomposition can be used to extract gene groups that seem to be specifically differently reported by different software, we next investigated whether a common feature in these gene groups can explain the differences.

3.4.2 Discordant alignment of reads on gene-pseudogene pairs

Expression mode 2 (EM2) is composed of genes that differ in quantification in regard to the alignment software used in the analysis. Based on the KNN score of Figure 2B, these genes are similarly quantified when using either HISAT2 or STAR, but differently when using TopHat2. In Figure 3A, we can observe the EM2 weight distribution for all genes, where the genes considered significant (more than 4 standard deviations from the mean) are colored in blue and red, for positive and negative weights, respectively. In Supplementary Figure 2, we can observe the distribution of the pipelines along EM2. Seeing that HISAT2 and STAR have bigger projection scores than TopHat2, we can infer that the genes with

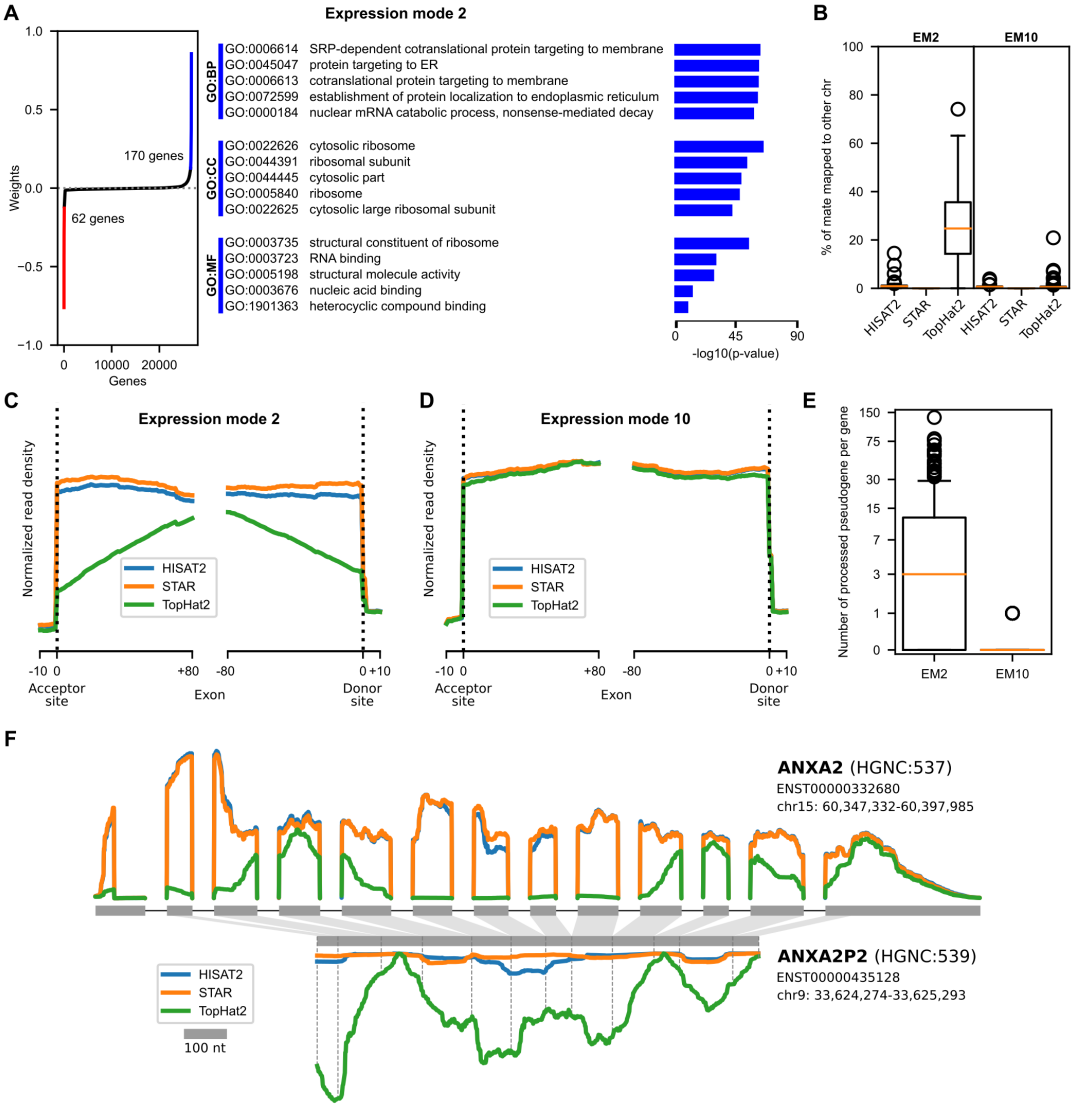


Figure 3 – Description of the features explaining the alignment software classification. **A** illustrates the gene weights distribution of EM2, along with their Gene Ontology (GO) enrichment analysis. Genes with a weight that is at least four standard deviations from the mean were considered significant and were coloured in the distribution. The GO enrichment analysis was performed using only the significant genes, and no enrichment was found in the negative genes. **B** shows the percentage of reads with mates that have been aligned onto another chromosome for the significant genes in EM2 and EM10, which is the negative control. The circles are the outliers of the boxplot. **C** is a metagene plot of the acceptor and donor sites of exon-exon junctions in EM2, with the profiles being separated by alignment software and **D** is generated using the EM10 genes. **E** quantifies the number of processed pseudogenes originating from the significant genes in EM2 in comparison to those in EM10. The plot is scaled using the inverse hyperbolic sine transformation. **F** describes the read profiles for each aligner along ANXA2 and ANXA2P2, the latter being the pseudogene of the former. The profiles are the averaged profiles of each considered pipeline and tissue combination. The exons are scaled accordingly to the 100 nucleotides (nt) reference in the legend. The introns were all truncated to a fix length in order to enhance readability. The mapping of the exons onto the pseudogenes was done using local sequence alignment, and position of the original splicing junctions are marked with dashed lines on the pseudogene. The pseudogene read profile is presented upside-down, and the two read profile plots are presented using the same scale.

a positive weight are more highly expressed when using STAR and HISAT2 than when using TopHat2, and vice versa.

Figure 3A also displays results from a Gene Ontology (GO) enrichment analysis of the significant genes. This analysis was done separately for the genes with positive and negative weights. Only results of the enrichment for the positive genes are shown since the analysis of the negative genes led to no significant enrichment. Interestingly, we found a strong enrichment for ribosome and translation related GO terms in the positive genes group. For this to happen, the common feature of the positive genes that is considered differently by the alignment software must also be linked to some biological characteristics of the genes. By exploring the average alignment statistics for the positive genes in EM2, we found that TopHat2 has a significant percentage of mapped read pairs with a mate aligned to another chromosome, as displayed in Figure 3B. For the same genes, STAR reports no read pairs mapped to different chromosomes, and HISAT2 has an average of 1% (in comparison to 25% for TopHat2) of mapped read pairs in this situation. In order to identify whether the observed effect is expression mode dependent, we used another unrelated expression mode as a control. By comparing these results to genes from a biological thyroid-related expression mode (EM10), we do not find the same effect. In EM10, the three software tools have a similar and a nearly null number of read pairs mapped to different chromosomes. Having identified a divergent characteristic, we then analyzed the discordant read pairs, by comparing alignments from STAR and HISAT2 to alignments from TopHat2.

Metagene plots in Figure 3C and 3D illustrate the aggregation of read profiles for all exon acceptor and donor sites for the gene groups of interests. While Figure 3D shows a very similar read profile for all alignment software, Figure 3C shows a dissimilar profile for TopHat2. Thus, HISAT2 and STAR profiles are similar in EM2 and EM10 but the TopHat2 profile in EM2 lacks reads around the exon-exon junctions. A theoretical exon-exon junction profile would show a perfectly square profile at the acceptor and donor sites. In our case, the progressively smaller profile, as we approach the edge of the exon, is a sign of a difficulty to align reads spanning across an exon-exon junction. Because we are aligning on the genome, alignment software must be able to map reads in a discontinuous manner across an exon-exon junction, namely gapped alignment. All three tested aligners are known to perform gapped alignment, but TopHat2 seems to fail to do so in the specific situation highlighted by the ICA expression mode. The biological particularity of the EM2 positive genes is that they possess a significantly higher number of processed pseudogenes in comparison to the other genes, as illustrated by Figure 3E. The gene-pseudogene rela-

tionship used is described by the PsiCube project [42]. Processed pseudogenes are defined as the product of the retrotranscription of spliced RNA inserted back into the genome. This means that they have the genomic sequence of the transcribed product of their parent gene, i.e. continuous exon-exon junction sequence. In our situation, TopHat2 prefers using a distant already spliced junction than using a local junction that needs splicing. The creation of processed pseudogenes has been shown to favor highly conserved genes that are widely expressed [43], without any detectable sequence bias [44]. This definition fits well with the ribosomal proteins and translation associated machinery found in the GO enrichment analysis.

3.4.3 Expression modes each identify gene groups with opposite behaviours

Having demonstrated that the positive genes in EM2 have lower read mapping in TopHat2 due to the presence of pseudogenes, we turned our attention back to EM2 negative genes. Due to the two-tailed distribution of gene weights, we would expect an opposite effect, meaning that the negative genes should have a higher read mapping in TopHat2 than HISAT2 and STAR. We should also point out the abnormal (when compared to the other weights distributions) shape of the negative gene weights, harboring a very steep change of weight, instead of the progressive asymptotic-like shape of the distribution. Using Ensembl 98 biotypes, we found that the majority of the negative genes are pseudogenes (44/62) and the remainder are protein-coding genes (18/62) and those are primarily mono-exonic (13/18). In Figure 3F, we illustrate a pair of genes, AXNA2 and AXNA2P2, that were both found in EM2, both in opposite gene groups. AXNA2P2, as its symbol indicates, is a pseudogene originating from AXNA2, and the correspondence between the two, established with local alignment of their mature RNA sequences, shows that the pseudogene is a truncated intron-less copy of a transcript from the original gene. Averaged read profiles from all considered RNA-seq pipelines, separated by alignment software, are shown for both genes. TopHat2 profiles are obviously quite different from the other aligners, but more interestingly, both of its profiles seem to overlap, where the sum of the two profiles is similar to HISAT2 and STAR profiles of the principal gene. For example, the peak of TopHat2 reads on the ANXA2 fourth exon is related to the minimum value found in the ANXA2P2 corresponding section. This figure provides additional proof that TopHat2 shares the reads between a gene and its pseudogenes, while HISAT2 and STAR do not. Supplementary Figure 3 provides three other pairs of genes illustrating the same situation. To further our argument, we can observe that the exon length is responsible for part of the

TopHat2 profile. In the RPL13A profile, the exons are too short for reads to be mapped exclusively onto one exon, creating a situation where the pseudogene is getting nearly all the reads. Conversely, GLUD1 has longer exons, and we can observe the same exon profile as found in the metagene plot Figure 3C. GLUD1 also shows, looking at its last exon, that as soon as we are in an exon long enough to map entire reads, the three aligner profiles converge. We believe that the steep change observed in the weight distribution originates from having a binary feature which is being a product of a spliced retrotranscription or not.

3.4.4 Genome annotations and quantifiers interact in RNA-seq quantification

Expression modes 3, 5 and 16 offer a more convoluted story, because both the genome annotations and the quantification software have been partly clustered in them. This means that these two steps, for the genome annotations and software selected in this study, are not fully independent. In Supplementary Figure 4, we can observe the distributions of the different pipeline projections along the three technical modes of interest, colored by their annotation and quantification software. The same pattern can be seen in the three technical modes. In both steps, we can find two different, loosely defined but clearly separated, clusters. One cluster always contains two genome annotations and the same two quantifiers, featureCounts and HTSeq, while the other cluster contains the three genome annotations and the three quantifiers. Because quantification is downstream of the genome annotation specification, we will illustrate this phenomenon as being quantification driven. The hypothesis here is that there are some features in the genome annotations upon which a classification can be done if the pipelines used featureCounts or HTSeq. These different features can also be used to generate all three possible binary classification of annotations. However, when using Cufflinks as the quantification software, these features appear to have no effect on the quantification, because Cufflinks is shown in a single cluster. Therefore, Cufflinks seems to rescue the problematic features which cause quantification biases when using featureCounts or HTSeq.

3.4.5 Expression modes may hide other gene groups with similar quantification power

This situation lets us test a hypothesis that we spelled out earlier, which is the fact that some expression modes can hide other, smaller, expression modes. To test this, and to also put Cufflinks to the test, we generated another ICA model, in which we only used

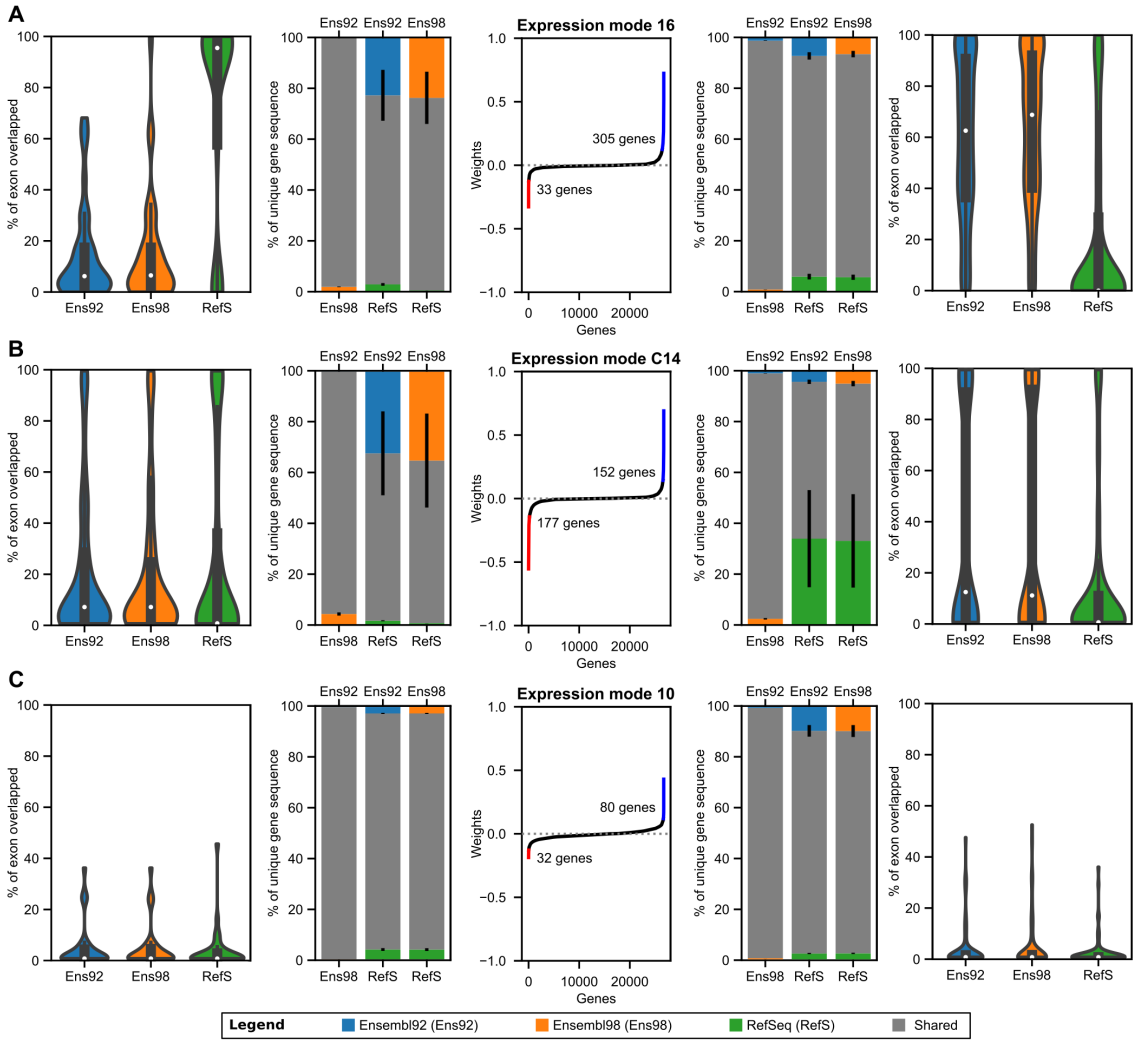


Figure 4 – Description of the features leading to an Ensembl versus RefSeq classification in our expression datasets. **A** and **B** are technical modes linked with genome annotation classification and **C** is a negative control, a biological mode linked with thyroid-related gene groups. The middle plots describe the distribution of gene weights in the mode, highlighting the genes considered to be significant (more than 4 standard deviations from the mean). On both sides of the middle plot, sets of similar plots are found. The plots on the left show metrics calculated from the significant genes with a negative score, and conversely for the right plots. The outer most plots show the distribution of the percentage of a gene exons, by nucleotide length, that are overlapped by exons from other genes. Overlapping genes from the sense and antisense strands were both considered. The inner flanking plots show the average of shared and unique exon sequences for each pairwise combination of genome annotations. Sections drawn in an annotation color represent the percentage of sequence not found in the other annotation, while the shared section can be found in both annotations. The black lines represent the error bars. For both flanking plots, each individual genomic position was only considered once, independently of the number of individual exons of the same gene that may overlap that position.

expression datasets that were generated using Cufflinks. This ICA model was processed in the same way as previously described, with M = 16 being the model with the smallest MSE and the highest number of expression modes. The KNN score heatmap of this ICA model can be found as Supplementary Figure 5A. In this Cufflinks-only ICA model, we can find two different technical modes, independently related to the annotation and the aligner blocks. Both modes seem to have the same classification power (KNN score pattern over the different tools in a block) as another technical mode found in the original ICA model. To verify whether we have found the same expression modes in two different models, or expression modes based on different gene groups displaying the same classification power, we can compare the overlap of significant genes in both expression modes. This overlap is shown in Supplementary Figure 5B, where M2 is compared to MC7, both being similar alignment technical modes, and where M16 is compared to MC14, both being similar annotation technical modes and both able to separate Ensembl from RefSeq. In the first case, we can observe that the genes from the two alignment technical modes largely overlap, which means that they have probably learned from the same gene groups, using the same features. Conversely, the technical modes describing the annotations have a small overlap, meaning that the two modes have been built on mainly independent gene groups, which also means different features. The overlap might be explained by genes having features that let them be part of both groups. We therefore have two different expression modes, dependent on the quantification software used, that have classification power over RefSeq and Ensembl genome annotations.

3.4.6 A gene quantification is affected by its definition and its neighbouring gene definitions

Figure 4 explains the main features used by expression modes 16 and C14 to drive the clustering. Expression mode 10, the thyroid-related biological mode, was also used as a negative control, because it is not expected to be enriched for the annotation classifying features. First, we looked at the extent of exon overlap for the genes considered, within each annotation. To do so, we measured the proportion of exonic nucleotides that overlap any other gene in the given annotation, considering both sense and antisense strands (outer most plots in Figure 4). These analyses indicate that negative and positive significant genes in M16 are exhibiting an opposite pattern which is not found in MC14 and M10. The negative M16 genes are highly overlapped in RefSeq while being marginally overlapped in Ensembl 92 and 98, and inversely for the positive genes. Next we compared the annotations pairwise and measured the percentage of unique and common sequences for the genes of

interest when two annotations are compared (flanking inner plots). These analyses show that MC14 is now also exhibiting a mirroring pattern, where its negative significant genes are longer and have more unique sequence in Ensembl than RefSeq, while the opposite is true for its positive genes. M10 shows that some baseline of sequence differences might be expected, with Ensembl having more unique sequences than RefSeq, but only C14 has a clearly defined mirror effect. This brings us to the observation that, because the ICA generates two-tailed distributions for the gene weights, we are able to see genes exhibiting both extremes of a feature, as clearly shown by the mirroring effect in Figure 4. While it is clear that both technical modes have a main feature, where M16 is mainly affected by overlapping annotations and MC14 by different gene definitions, we can also see aspects of the other feature in both sections. This might be due to the fact that the two technical modes share some of their genes, leading to genes contributing in both sections. One can also easily imagine how having a longer gene sequence might also result into having more overlapping sequences, showing that these two features are not completely independent.

3.4.7 Extrinsic and intrinsic factors of genome annotations affect quantification differently

Interestingly, the quantification software behaviour classifies them by the approach that they use to solve the quantification problem. Both featureCounts and HTSeq are count-based quantifiers, whereas Cufflinks quantifies through transcript assembly and quantification. We have observed that the main bias of count-based quantification is the definition of neighbouring genes, seen through the metric describing the percentage of exon overlap, whereas transcript quantification is primarily affected by the definition of the gene itself. We will describe the former bias as originating from the extrinsic factors, and the latter from the intrinsic factors. By projecting expression datasets from featureCounts and HTSeq onto the technical mode C14, as seen in Supplementary Figure 5C, we can see that these two tools are also being affected by the intrinsic factors, because they are shown to cluster in the same way as Cufflinks does. The fact that MC14 does not appear in the original ICA model means that M16, the extrinsic factors affecting featureCounts and HTSeq, exhibits a stronger bias that is hiding the bias caused by intrinsic factors. We can also pose that the intrinsic factors are independent, meaning that all quantification software behave the same way towards them, whereas the extrinsic factors are the ones responsible for the interaction between genome annotations and quantifiers, because they are treated differently by the software.

Supplementary Figure 6 illustrates three groups of example genes that were found to be significant in at least one of the two technical modes classifying Ensembl and RefSeq. All three genes of interest are protein-coding genes with conserved consensus coding sequence [45] (CCDS) across all three genome annotations. The first gene, ARPC1A, was found in EM16 and is a good example of a gene being overlapped by a read-through gene that is unknown from RefSeq. The second example, GOLGA8M, was found in EMC14 and is shown to have major differences in annotation from Ensembl to RefSeq. There are also some overlapping genes that are not the same, but they are smaller in comparison to the overall gene. Intriguingly, Ensembl possesses two different genes (ENSG00000188626 and ENSG00000261480) that have the same gene symbol and reported HGNC ID, while HGNC only lists the first as being GOLGA8M. Moreover, the first has two transcripts which are named GOLGA8M-201 and GOLGA8M-202, while the second gene only has GOLGA8M-203. The last example, KCNA6, was found in both EM16 and EMC14 and exhibits both a change in overlap and a change in definition across Ensembl and RefSeq. This figure provides a better understanding of the separation between extrinsic and intrinsic factors.

3.4.8 ICA can correctly classify different versions of the Ensembl genome annotation

We have shown that it is possible to classify RNA-seq quantification results with respect to their source of genome annotation. To delve deeper into this issue, we have also included in the study two different versions of the Ensembl genome annotation, versions 92 and 98, which were respectively published in April 2018 and September 2019, approximately 18 months apart. We wanted to see whether we could also classify them, and if applicable, identify the way they diverge from one another. The ICA model has identified two technical modes, EM3 and EM5, that differentiate between the two Ensembl versions, with RefSeq pairing with a different Ensembl version in each technical mode, as seen in Supplementary Figure 4. Once again, Cufflinks does not behave as the two other quantifiers, affecting the global clustering of the annotation projections. From our previous observations on EM16, we can hypothesize that EM3 and EM5 classifications will also be driven by extrinsic factors, meaning genes with overlapping loci. Supplementary Figure 7 presents the data supporting our explanations of the clustering, with the same plot representations as used with the previous technical modes. However, both technical modes do not offer the same prominent mirroring effect as observed in Figure 4. If these modes are also driven by extrinsic factors, it would be expected that the outer plots show a clear difference in the

percentage of exon overlap. In EM5 (Supplementary Figure 7B), positive genes display a clear separation between Ensembl 98 and the other annotations when quantifying the percentage of exon overlap. Notably, the distribution of scores for the two other annotations is approximately the same, as it was for the same plots in Figure 4. With the mirroring hypothesis, we would expect that Ensembl 98 would exhibit a smaller percentage of overlapped exon in the negative genes, which it does. However, its distribution is not clearly different from Ensembl 92, and Ensembl 92 and RefSeq distributions do not look alike. Both of these points show a divergence from the previously observed data. In order to explain this discrepancy, we characterized the difference in percentage of exon overlap for each gene in the components, across the genome annotations. Supplementary Figure 7C shows this characterization for both EM3 and EM5, the plots being relative to the annotation that is clustered alone, which is respectively Ensembl 92 and Ensembl 98. Positive and negative genes are split into three groups, with respect to their difference in exon overlap with the reference annotation. If both scores of a gene are within 10% of the reference annotation score, the gene will be put in the middle group, being approximately the same as the reference annotation. If the gene has a score of at least 10% higher or lower in at least one annotation, it will be classified accordingly in the bigger than or smaller than group, while also being colored relative to which, or both, annotation diverges from the base case. If both scores are outside the 10% threshold in opposite directions, the gene will be classified as similar to the reference annotation. Looking at the results for EM5, we can see a big difference between the positive and negative genes. Based on the mirroring, it was expected that the genes would be separated mainly in a diagonal fashion, but the new information that we gain from this representation is the unequal separation of the exon overlap score across the two annotations. While the positive genes are mainly below the threshold for both annotations, the negative gene group is dominated by genes that are only differing in RefSeq. The same kind of grouping can be observed in EM3, where the majority of the genes contributing to the exon overlap score are only positive for one annotation. The positive genes plot displaying percentage of unique gene sequence in EM3 also shows that there might be some influence of intrinsic factors as well, with Ensembl 98 and RefSeq having more unique sequence compared to Ensembl 92. Finding large enough differences to enable classification of two different genome annotation versions is a much more difficult task than for two different annotations, and the technical modes demonstrate this by using different sets of genes to be able to cluster differently. We believe that these gene groups are heterogeneous groups, with each subpart contributing to differentiate with one annotation at a time.

3.4.9 Ensembl distinguishes itself by a higher, and growing, number of overlapping loci

EM3 and EM5 let us explore the differences in Ensembl, and posit in the way that Ensembl is currently evolving through time. Interestingly, the clarity of the gene groups observed in Supplementary Figure 7C may be used to interpret the evolution of the annotations. Based on the projections, Ensembl 98 is less expressed in the positive genes, and to achieve that, we only need to find genes that have gained overlapping loci in Ensembl 98 and that are not overlapped in RefSeq. Conversely, to be more expressed in EM5 negative genes, Ensembl 98 needs to lose an overlapping loci from Ensembl 92, loci that must be present in RefSeq. From the results, we can conclude that it is easier for Ensembl 98 to gain new overlapping annotation than to lose some from Ensembl 92, and we can also conclude that RefSeq seems to globally possess fewer overlapping genes, because it is mainly responsible for the EM5 negative genes. In order to explore this, we quantified the percentage of overlapped exon across all of the 26 713 genes considered in this study. This quantification is shown in Supplementary Figure 7D, as distributions only including genes that have a non-null overlap. Quite surprisingly, the distributions for the different genome annotations are essentially that same, made from a different number of genes. While anecdotally looking at the genes responsible for the new overlaps in Ensembl 98, we stumbled upon many read-through, that were not necessarily identified as such, and set forth to quantify them in Supplementary Figure 7E. We defined a read-through as a gene having at least a transcript that has, for a least two different genes, a perfectly matching exon, based on the genomic coordinates, and not the sequence alone. These matching exons must also be distinct exons in the overlapping transcript. We also only quantified read-throughs that are overlapping at least one gene included in our study. When comparing Ensembl 92 and RefSeq, based on the last two plots, we can acknowledge that Ensembl has far more genes with overlapped loci (38% against 27% for RefSeq), and about three times more read-throughs than RefSeq. We can also see that Ensembl 98 is straying further away from RefSeq, with Ensembl 98 having even more overlapped genes and read-through genes.

3.4.10 Differential expression analyses describe the same extent of methodological step biases

A differential expression analysis (DEA) is usually used to identify whether a gene expression is varying accordingly to an experimental variable [46]. In comparison to the ICA analysis that we performed, DEA reports all genes independently, whereas gene groups

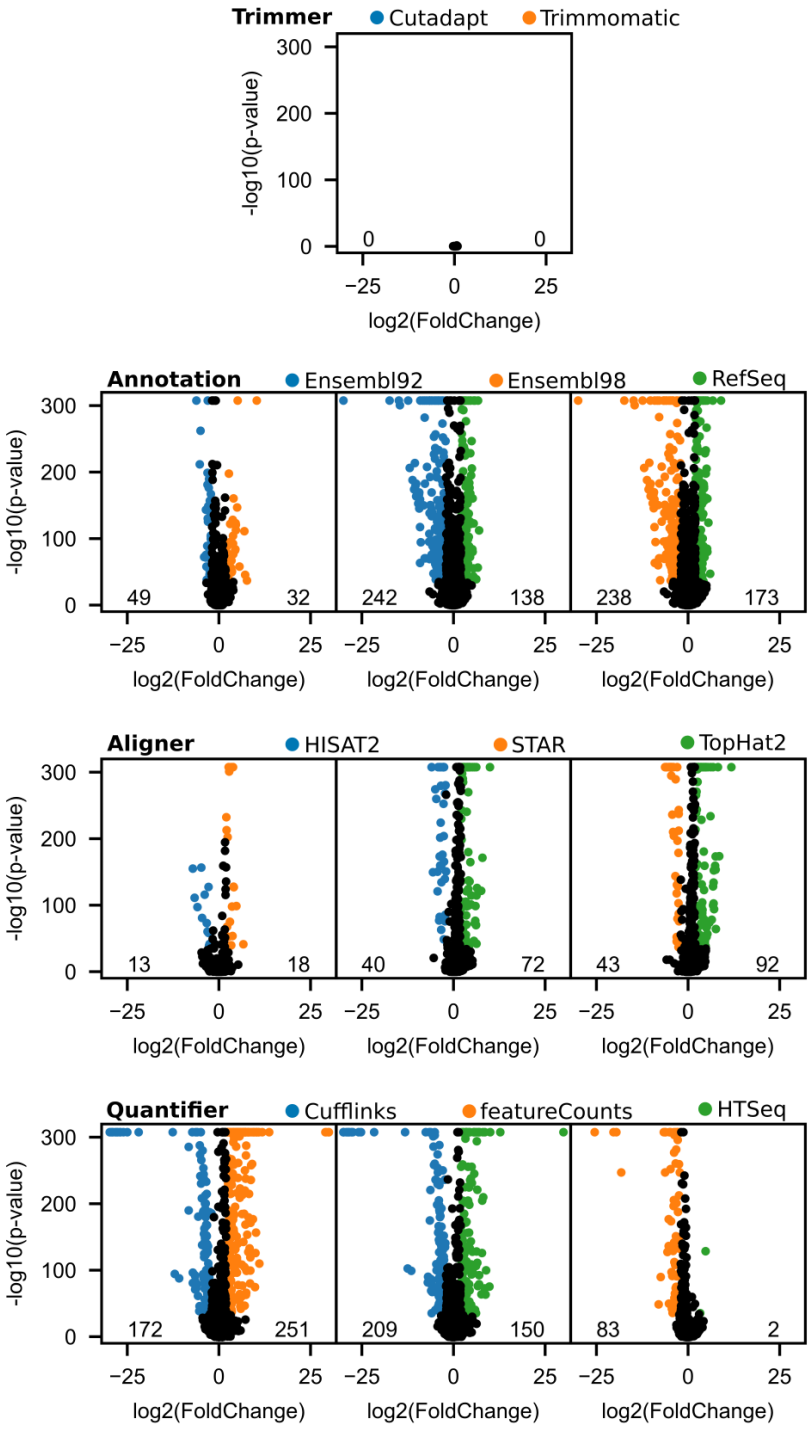


Figure 5 – A differential expression analysis was performed for each pairwise combination of choices for each pipeline step. The volcano plots of these analyses are presented here, where significantly differentially expressed genes (p-value < 10e-35 and log2(fold change) >= 2) are colored accordingly to the choice in which it is overexpressed. Numbers on the lower edge of the plots quantify the number of colored genes.

in the ICA have some shared expression patterns. In order to compare the genes found through the ICA to a more commonly used technique in the RNA-seq field, we performed several DEA using the pipeline variables and their choices as respectively experiments and conditions. For example, to generate a DEA on the impact of the trimming, we compared expression datasets processed using Trimmomatic to those processed using Cutadapt, using all the corresponding datasets as replicates. This means that our replicates are very heterogeneous, having expression datasets grouped together that were generated using the whole spectrum of the other pipeline steps. In a DEA, negative results do not mean that the genes were not impacted by the condition, but that such an impact could not be observed within the variance of the datasets. Because some pipelines do have a significant impact on the quantification, other steps might suffer from large within-group variance, which makes them less likely to have significant results. Our multiple DEA analyses have very heterogeneous within-group variance due to the fact that we are reanalyzing the same expression datasets, with different groupings of the samples with respect to the different methodological steps.

The number of replicates is usually a variable worth optimizing in DEA, since more replicates means more sequencing, and sequencing is still a costly task [47]. In our case, the replicates are mainly generated through processing the same dataset using a different in silico pipeline, meaning that we have an abnormally large number of replicates for the different experiments. These replicates translate into abnormally small p-value, and we even hit the number limit of a 64-bit system, where number smaller than approximately 10e-308 are considered as 0 and reported by DESeq2 as such [34]. In the volcano plots, any gene with a p-value of 0 was displayed as having (in -log10 form) the maximum possible value.

Figure 5 displays volcano plots for the different technical DEA, grouped by pipeline step. Using a p-value of 10e-35 and a log2(fold change) of 2 as our significance thresholds, we have identified the significantly differently expressed genes (DEGs) for the different experimental choices using their colors and displayed their counts on the lower edge of the volcano plots. It is apparent from an overview of the different pipeline steps that the methodological choices do not bear the same importance in the definition of the results. We can also immediately see some resemblance with the ICA results. The trimming step is the only step not highlighted by the ICA, and all of its genes in the volcano plot are centered around the null coordinates, nowhere near significance. Accordingly, the alignment step was captured by only one technical mode in the ICA model, and has the second-lowest

number of DEGs. It can also be observed that HISAT2 and STAR generate results that are more similar than TopHat2. And lastly, genome annotations and quantifiers seem to have a similar number of DEGs, and the same tool clustering as found in the ICA is observable.

3.5 Discussion

To our knowledge, the community has overlooked the "what-to-benchmark" question regarding the RNA-seq quantification pipeline [48]. This question might be more akin to a manufacturing context, where optimization of resources is directly linked to the metric of success, whereas it can be more fuzzily linked in research. The "what-to-benchmark" question is in fact of utmost importance in any resource-limited context, such as research, to correctly identify the most pressing questions. In the context of RNA-seq, the "what-to-benchmark" question is related to the action of benchmarking the sensitivity of the pipeline as a whole, in order to identify the greatest source of variability in the results.

In this study, we described quantification biases found in the whole RNA-seq in silico workflow, using an ICA decomposition on expression datasets created through reprocessing of RNA-seq experiments by a wide variety of methodological choices. We put forward the idea of technical reprocessing to identify biases between different choices of a specific methodological step, instead of assessing the deviation from the ground truth, which is a difficult value to find. Biases in a specific methodological step are proof that developers either used different biological hypotheses or have encountered unknowns. While this approach is not a panacea, since it will fail to identify any problematic shared by all the methodological choices, it can be an interesting technique to characterize the current landscape of quantification discrepancies concerning tools presently used in the literature. Such landscape can be characterized by a within step manner, which is the usual analytical benchmarking approach, but also a between step manner, which highlights the relative importance of the different methodological choices. The idea of gene specific biases identification in a methodological context in not new, but we extent this approach by being able to link similarly affected genes together, leading to an easier identification of impactful characteristics of problematic genes [7].

While the format of this study does not let us provide direct recommendations regarding what methodological choices to prefer, we can highlight choices that we would not recommend. While Tophat2 is still one of the primary aligner software in use in the lite-

rature [17], we cannot recommend its usage due to its bias towards genes with processed pseudogenes. Furthermore, HISAT2 is a reimplementation of TopHat2, correcting some issues that have become apparent through the years. Authors of TopHat2 have also published many times, as seen in TopHat2 and HISAT manuals, and on the twittersphere, about the need to move towards newer software. While further models might highlight discrepancies between STAR and HISAT2, the data shown here cannot point towards one or the other.

We cannot recommend using count-based quantifiers, such as featureCounts and HTSeq, since, as stated by themselves [30], they are not able to give accurate estimation of isoform quantification. The issue is that the isoform quantification problem is actually an overlapping transcripts problem, which means that it is not limited to transcripts of a single gene. If no genes were overlapping, every read falling into the region of a certain gene could be trivially assigned to it. But, as demonstrated by the technical modes linked to the quantification software, gene-level quantification for count-based quantifiers fails when the genes are overlapped by other genes. This also means that count-based quantifiers have overlooked the fact that genes may share some genomic coordinates. Based on that information, it would make sense to recommend using quantifiers that quantify on the transcript level through techniques trying to infer the transcripts from which each read could have been produced, a class of quantifiers represented by Cufflinks in this study [49]. As we have demonstrated, such software seems to be less affected by genome annotation extrinsic factors.

As for the choice of genome annotations, the recommendations are not as clear. Genome annotations do not hold the same place as the other software tools in the RNA-seq pipeline. Performance of software can be tested under specific conditions, and divergence from the expected behavior can be assessed. This was highlighted by technical modes, for example the lack of reads on exon-exon junctions for some genes when using TopHat2 contradicts our understanding of gene expression and the hypothesis of uniform distribution of reads along a gene. But genome annotations are information resources, having the dual purpose of being a repository of our gene biology knowledge, and a research tool, leveraging high-throughput techniques. While the technical modes associated with genome annotations informed us of differences in the annotation of different genes, differences that have a variety of impact on the basis of the quantification software used, it is difficult to assess which genome annotation is closer to the truth. Learning about differences that are driving the main biases in quantification is important for our understanding of the place of genome annotations in the RNA-seq pipeline, and while it does not give a clear answer

about which annotation to use, it informs users about the non-triviality of choosing an annotation, and about features that are important to look for. On the other side, it is not that the annotations have converged on a similar understanding of the structure of a gene that this convergence is in phase with our current understanding of biology. To illustrate this, we can take the extreme example of the GDF1 and CERS1 genes that individually have an identical gene-level structure across Ensembl92, Ensembl98 and RefSeq, while also sharing the vast majority of their exons, as seen in Supplementary Figure 8. Overlapping loci are difficult to quantify, which makes these genes susceptible to unreliable quantification across the methodological landscape, but it also raises the question of whether these two entities are really independent genes. GDF1 and CERS1 even share an identical CDS in Ensembl through two transcripts (GDF1-201 and CERS1-205) that only differs by some nucleotides in both UTR extremities. It was proposed that GDF1 produces a polycistronic mRNA [50], but the two proteins discussed in the paper are now products of the two different genes. Genome annotations are not a data structure that is currently able to support polycistronic RNA, and it might be what justified the separation into two different genes. But knowing that overlapping genes cause issues in quantification, and that, if they are truly polycistronic, there is only one RNA to quantify, we should review the genome annotation information structure. There have been multiple projects exploring the polycistronic nature of human mRNAs (e.g. [51, 52]). The inclusion of such data would require allowing transcripts to possess multiple CDS elements. This is simply an example that biological understanding and hypotheses evolve, and that our software, and usage of them, must follow accordingly.

To compare the genome annotations, we had to limit ourselves to genes that have some basis of comparison, and HGNC was used to bridge the gene identification between RefSeq and Ensembl. Since not all genes are annotated into HGNC, and since not all HGNC genes have an identifier for both genome annotations, our number of comparable genes is smaller than the total number of genes, and one could hypothesize that the remaining genes are better described and are probably more similar than the other genes. Furthermore, some of the remaining genes were not reported by the pipelines when using RefSeq. These genes, the 37 mitochondrial genes and 10174 pseudogenes (list available as Supplementary Data 3), are present in the RefSeq GTF annotation file, but are described using a single ‘gene’ feature, instead of the expected hierarchy where each gene possesses one or more transcript, themselves having one or more exon. In the RefSeq GFF3 file, some of these genes appear to have the expected hierarchy, but they do not respect the same naming scheme as the other genes (NM and XM identifiers for transcripts), and they are always mono-exonic. This

issue has many implications. First, one would expect to have the exact same information within the different file formats distributed by a centralized resource. Second, deviating from the expected data format can create unexpected behavior for data processing software, seen here as genes not being quantified. Third, pseudogene is a very wide RNA category which includes mono-exonic processed pseudogenes, but also intron-bearing unprocessed pseudogenes. This makes for an unfair comparison if we were to compare quantifications of unprocessed pseudogenes between Ensembl and RefSeq. Supplementary Figure 9 provides such a visual comparison for the difference in gene structure. But because we do not have RefSeq quantifications for those genes, we cannot produce a quantitative comparison. Fourth, genes classified as pseudogenes in RefSeq may not be pseudogenes in all annotations. As described in Supplementary Data 3, Ensembl considers some of the RefSeq pseudogenes as having a different biotype, such as protein-coding, snRNA, lncRNA. This means that these genes might have had a proper annotation, and not simply a gene start and end coordinates, if RefSeq would be to consider them as a different biotype.

The differences between Ensembl versions also highlight the important question of what needs to be annotated ? Adding information to an annotation can have an impact on already existing annotations, as seen here and as previously described [53]. As an example, Ensembl is seen as having more read-through transcripts than RefSeq, and this number is getting bigger. Some of these transcripts have been shown to only be expressed in a cancer cell line, such as the HHLA1-OC90 read-through transcript in teratocarcinoma [54], while very few are actually found to be expressed more than anecdotally in non-cancerous cells [55]. As soon as a read-through annotation is added, count-based quantifiers will redistribute the counts of the overlapping genes. This should prompt users to move towards more appropriate software, but it also raises questions about the suitability of a ubiquitously used genome annotation, where annotation of rare events might affect the analysis.

Our ICA approach to identify gene group specific biases has proven itself useful, but must be used thoroughly and interpreted by a knowledgeable user. Removing quantification software from the model has produced a technical mode that was unseen in the first model. This means that generating ICA models using exhaustive combinations of inclusion and exclusion of software might reveal more technical biases that were hidden by present technical modes. Including a broader diversity of software would probably also deliver a larger diversity of technical modes. Transcriptome-based software and pseudo-aligners, despite the added difficulty of the combined steps, should be studied considering their growing place in the literature.

We believe that our approach contributes in answering the “what-to-benchmark” question. Our study provided data supporting the concept that the choice of a genome annotation plays an important role in gene quantification. This bias, like the others observed through the ICA for several pipeline steps, is not global, but affects specific gene groups sharing common features. We must emphasize the genome annotations because we believe that, in opposition to alignment and quantification tools, they have not received an appropriate amount of interest with respect to their importance in the definition of the results.

3.6 Acknowledgements

The authors wish to thank members of their groups for insightful discussions. The authors acknowledge Compute Canada as an outstanding resource for Canadian researchers. This work was supported by the Natural Sciences and Engineering Research Council of Canada [NSERC grant RGPIN-2018-05412 to MSS]. JS was supported by Masters scholarships from FRQNT and NSERC. MSS holds a Fonds de Recherche du Québec – Santé (FRQS) Research Scholar Junior 2 Career Award.

3.7 References

[1] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq : a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, jan 2009.

[2] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, no. 1, pp. 1–19, 2016.

[3] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling," *Science*, vol. 324, no. 5924, pp. 218–223, 2009.

[4] L. M. Weber, W. Saelens, R. Cannoodt, C. Soneson, A. Hapfelmeier, P. P. Gardner, A. L. Boulesteix, Y. Saeys, and M. D. Robinson, "Essential guidelines for computational method benchmarking," *Genome Biology*, vol. 20, no. 1, pp. 1–12, 2019.

[5] S. Ballouz, A. Dobin, T. R. Gingeras, and J. Gillis, "The fractured landscape of RNA-seq alignment : the default in our STARs.," *Nucleic acids research*, vol. 46, no. 10, pp. 5125–5138, 2018.

[6] N. A. Fonseca, J. Marioni, and A. Brazma, "RNA-Seq gene profiling–a systematic empirical comparison.," *PloS one*, vol. 9, no. 9, p. e107026, 2014.

[7] C. Robert and M. Watson, "Errors in RNA-Seq quantification affect genes of relevance to human disease," *Genome Biology*, vol. 16, no. 1, pp. 1–16, 2015.

[8] P. L. Germain, A. Vitriolo, A. Adamo, P. Laise, V. Das, and G. Testa, "RNAontheBENCH : Computational and empirical resources for benchmarking RNAseq quantification and differential expression methods," *Nucleic Acids Research*, vol. 44, no. 11, pp. 5054–5067, 2016.

[9] C. Everaert, M. Luypaert, J. L. Maag, Q. X. Cheng, M. E. DInger, J. Hellemans, and P. Mestdagh, "Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data," *Scientific Reports*, vol. 7, no. 1, p. 1559, 2017.

[10] S. M. E. Sahraeian, M. Mohiyuddin, R. Sebra, H. Tilgner, P. T. Afshar, K. F. Au, N. Bani Asadi, M. B. Gerstein, W. H. Wong, M. P. Snyder, E. Schadt, and H. Y. K. Lam, "Gaining comprehensive biological insight into the transcriptome by performing

a broad-spectrum RNA-seq analysis.," *Nature communications*, vol. 8, no. 1, p. 59, 2017.

[11] C. R. Williams, A. Baccarella, J. Z. Parrish, and C. C. Kim, "Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–13, 2017.

[12] C. Zhang, B. Zhang, L.-L. Lin, and S. Zhao, "Evaluation and comparison of computational tools for RNA-seq isoform quantification," *BMC Genomics*, vol. 18, p. 583, dec 2017.

[13] J. Costa-Silva, D. Domingues, and F. M. Lopes, "RNA-Seq differential expression analysis : An extended review and a software tool," *PLoS ONE*, vol. 12, no. 12, pp. 1–18, 2017.

[14] A. Baccarella, C. R. Williams, J. Z. Parrish, and C. C. Kim, "Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance," *BMC Bioinformatics*, vol. 19, p. 423, dec 2018.

[15] G. A. Merino, A. Conesa, and E. A. Fernández, "A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies," *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 471–481, 2019.

[16] J. Simoneau and M. S. Scott, "In silico analysis of RNA-seq requires a more complete description of methodology," *Nature Reviews Molecular Cell Biology*, vol. 20, pp. 451–452, aug 2019.

[17] J. Simoneau, S. Dumontier, R. Gosselin, and M. S. Scott, "Current RNA-seq methodology reporting limits reproducibility," *Briefings in Bioinformatics*, dec 2019.

[18] J. V. Stone, *Independent Component Analysis : A Tutorial Introduction*. Cambridge, Massachusett : MIT Press, 2004.

[19] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, pp. 51–60, jan 2002.

[20] N. Sompairac, P. V. Nazarov, U. Czerwinska, L. Cantini, A. Biton, A. Molkenov, Z. Zhumadilov, E. Barillot, F. Radvanyi, A. Gorban, U. Kairov, and A. Zinovyev, "Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets," *International Journal of Molecular Sciences*, vol. 20, p. 4414, sep 2019.

[21] A. Hyvärinen and E. Oja, "Independent component analysis : algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, jun 2000.

[22] E. Renard, S. Branders, and P. A. Absil, "Independent component analysis to remove batch effects from merged microarray datasets," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9838 LNCS, pp. 281–292, Springer, Cham, 2016.

[23] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, pp. 1260419–1260419, 2015.

[24] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, p. 10, may 2011.

[25] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic : a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, pp. 2114–2120, aug 2014.

[26] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2 : accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, p. R36, apr 2013.

[27] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nature Biotechnology*, vol. 37, pp. 907–915, aug 2019.

[28] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR : ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, jan 2013.

[29] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, pp. 562–578, mar 2012.

[30] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts : an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30,

pp. 923–930, apr 2014.

[31] S. Anders, P. T. Pyl, and W. Huber, "HTSeq–a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, pp. 166–169, jan 2015.

[32] F. Cunningham, P. Achuthan, W. Akanni, J. Allen, M. R. Amode, I. M. Armean, R. Bennett, J. Bhai, K. Billis, S. Boddu, C. Cummins, C. Davidson, K. J. Dodiya, A. Gall, C. G. Girón, L. Gil, T. Grego, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, J. C. Marugán, T. Maurel, A. C. McMahon, B. Moore, J. Morales, J. M. Mudge, M. Nuhn, D. Ogeh, A. Parker, A. Parton, M. Patricio, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, H. Sparrow, E. Stapleton, M. Szuba, K. Taylor, G. Threadgold, A. Thormann, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, A. D. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D745–D751, 2019.

[33] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (RefSeq) database at NCBI : Current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D733–D745, 2016.

[34] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, pp. 1–21, 2014.

[35] R. Dale, B. Grüning, A. Sjödin, J. Rowe, B. A. Chapman, C. H. Tomkins-Tinch, R. Valieris, B. Batut, A. Caprez, T. Cokelaer, D. Yusuf, K. A. Beauchamp, K. Brinda, T. Wollmann, G. L. Corguillé, D. Ryan, A. Bretaudeau, Y. Hoogstrate, B. S. Pedersen, S. van Heeringen, M. Raden, S. Luna-Valero, N. Soranzo, M. D. Smet, G. V. Kuster, R. Kirchner, L. Pantano, Z. Charlop-Powers, K. Thornton, M. Mar-

tin, M. van den Beek, D. Maticzka, M. Miladi, S. Will, K. Gravouil, P. Unneberg, C. Brueffer, C. Blank, V. C. Piro, J. Wolff, T. Antao, S. Gladman, I. Shlyakhter, M. de Hollander, P. Mabon, W. Shen, J. Boekel, M. Holtgrewe, D. Bouvier, J. R. de Ruiter, J. Cabral, S. Choudhary, N. Harding, R. Kleinkauf, E. Enns, F. Eggenhofer, J. Brown, P. J. Cock, H. Timm, C. Thomas, X. O. Zhang, M. Chambers, N. Turaga, E. Seiler, C. Brislawn, E. Pruesse, J. Fallmann, J. Kelleher, H. Nguyen, L. Parsons, Z. Fang, E. B. Stovner, N. Stoler, S. Ye, I. Wohlers, R. Farouni, M. Freeberg, J. E. Johnson, M. Bargull, P. R. Kensche, T. H. Webster, J. M. Eppley, C. Stahl, A. S. Rose, A. Reynolds, L. B. Wang, X. Garnier, S. Dirmeier, M. Knudsen, J. Taylor, A. Srivastava, V. Rai, R. Agren, A. Junge, R. V. Guimera, A. Khan, S. Schmeier, G. He, L. Pinello, E. Hägglund, A. S. Mikheyev, J. Preussner, N. R. Waters, W. Li, J. Capellades, A. T. Chande, Y. Pirola, S. Hiltemann, M. L. Bendall, S. Singh, W. A. Dunn, A. Drouin, T. D. Domenico, I. de Bruijn, D. E. Larson, D. Chicco, E. Grassi, G. Gonnella, J. B, L. Wang, F. Giacomoni, E. Clarke, D. Blankenberg, C. Tran, R. Patro, S. Laurent, M. Gopez, B. Sennblad, J. A. Baaijens, P. Ewels, P. R. Wright, O. M. Enache, P. Roger, W. Dampier, D. Koppstein, U. K. Devisetty, T. Rausch, M. Cornwell, A. E. Salatino, J. Seiler, M. Jung, E. Kornobis, F. Cumbo, B. K. Stöcker, O. Moskalenko, D. R. Bogema, M. L. Workentine, S. J. Newhouse, F. d. V. Leprevost, K. Arvai, and J. Köster, "Bioconda : Sustainable and comprehensive software distribution for the life sciences," *Nature Methods*, vol. 15, no. 7, pp. 475–476, 2018.

[36] J. Köster and S. Rahmann, "Snakemake–a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, pp. 2520–2522, oct 2012.

[37] B. Yates, B. Braschi, K. A. Gray, R. L. Seal, S. Tweedie, and E. A. Bruford, "Genenames.org : The HGNC and VGNC resources in 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D619–D625, 2017.

[38] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, p. R106, oct 2010.

[39] A. Kessy, A. Lewin, and K. Strimmer, "Optimal Whitening and Decorrelation," *The American Statistician*, vol. 72, pp. 309–314, oct 2018.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn : Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[41] P. V. Nazarov, A. K. Wienecke-Baldacchino, A. Zinovyev, U. Czerwińska, A. Muller, D. Nashan, G. Dittmar, F. Azuaje, and S. Kreis, "Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients.," *BMC medical genomics*, vol. 12, p. 132, sep 2019.

[42] C. Sisu, B. Pei, J. Leng, A. Frankish, Y. Zhang, S. Balasubramanian, R. Harte, D. Wang, M. Rutenberg-Schoenberg, W. Clark, M. Diekhans, J. Rozowsky, T. Hubbard, J. Harrow, and M. B. Gerstein, "Comparative analysis of pseudogenes across three phyla," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 37, pp. 13361–13366, 2014.

[43] I. Gonçalves, L. Duret, and D. Mouchiroud, "Nature and structure of human genes that generate retropseudogenes.," *Genome research*, vol. 10, pp. 672–8, may 2000.

[44] C. Esnault, J. Maestre, and T. Heidmann, "Human LINE retrotransposons generate processed pseudogenes," *Nature Genetics*, vol. 24, pp. 363–367, apr 2000.

[45] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M.-M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. DiCuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, "The consensus coding sequence (CCDS) project : Identifying a common protein-coding gene set for the human and mouse genomes," *Genome Research*, vol. 19, pp. 1316–1323, 2009.

[46] D. K. Slonim, "From patterns to pathways : Gene expression data analysis comes of age," *Nature Genetics*, vol. 32, no. 4S, pp. 502–508, 2002.

[47] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton, "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use ?," *RNA (New York, N.Y.)*, vol. 22, pp. 839–51, jun 2016.

[48] F. Y. Partovi, "Determining What to Benchmark : An Analytic Hierarchy Process Approach," *International Journal of Operations & Production Management*, vol. 14,

no. 6, pp. 25–39, 1994.

[49] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, pp. 511–515, may 2010.

[50] S.-J. Lee, "Expression of growth/differentiation factor 1 in the nervous system : Conservation of a bicistronic structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 10, pp. 4250–4254, 1991.

[51] S. A. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, A. D. Karger, B. A. Budnik, J. L. Rinn, and A. Saghatelian, "Peptidomic discovery of short open reading frame-encoded peptides in human cells," *Nature Chemical Biology*, vol. 9, no. 1, pp. 59–64, 2013.

[52] M. A. Brunet, M. Brunelle, J. F. Lucier, V. Delcourt, M. Levesque, F. Grenier, S. Samandi, S. Leblanc, J. D. Aguilar, P. Dufour, J. F. Jacques, I. Fournier, A. Ouangraoua, M. S. Scott, F. M. Boisvert, and X. Roucou, "OpenProt : A more comprehensive guide to explore eukaryotic coding potential and proteomes," *Nucleic Acids Research*, vol. 47, no. D1, pp. D403–D410, 2019.

[53] A. B. Pyrkosz, H. Cheng, and C. T. Brown, "RNA-Seq Mapping Errors When Using Incomplete Reference Transcriptomes of Vertebrates," *arXiv*, mar 2013.

[54] P. E. Kowalski, J. D. Freeman, and D. L. Mager, "Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes," *Genomics*, vol. 57, no. 3, pp. 371–379, 1999.

[55] M. Babiceanu, F. Qin, Z. Xie, Y. Jia, K. Lopez, N. Janus, L. Facemire, S. Kumar, Y. Pang, Y. Qi, I. M. Lazar, and H. Li, "Recurrent chimeric fusion RNAs in non-cancer tissues and cells," *Nucleic Acids Research*, vol. 44, no. 6, pp. 2859–2872, 2016.

[56] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, p. R25, mar 2009.

[57] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, apr 2012.

[58] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, pp. 1754–1760, jul 2009.

[59] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 26, pp. 589–595, mar 2010.

[60] T. D. Wu and S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," *Bioinformatics*, vol. 26, pp. 873–881, apr 2010.

[61] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT : a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, pp. 357–360, apr 2015.

[62] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu, "MapSplice : Accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Research*, vol. 38, pp. e178–e178, oct 2010.

[63] J. Hu, H. Ge, M. Newman, and K. Liu, "OSA : a fast and accurate alignment tool for RNA-Seq," *Bioinformatics*, vol. 28, pp. 1933–1934, jul 2012.

[64] H. Jiang and W. H. Wong, "SeqMap : mapping massive amount of oligonucleotides to the genome," *Bioinformatics*, vol. 24, pp. 2395–2396, oct 2008.

[65] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat : discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, pp. 1105–1111, may 2009.

[66] A. R. Quinlan and I. M. Hall, "BEDTools : a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, mar 2010.

[67] P. Glaus, A. Honkela, and M. Rattray, "Identifying differentially expressed transcripts from RNA-seq data with biological variation," *Bioinformatics*, vol. 28, pp. 1721–1728, jul 2012.

[68] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data.," *Genome research*, vol. 22, pp. 2008–17, oct 2012.

[69] A. Roberts and L. Pachter, "Streaming fragment assignment for real-time analysis of sequencing experiments.," *Nature methods*, vol. 10, pp. 71–3, jan 2013.

[70] S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis, "Transcriptome genetics using second generation sequencing in a Caucasian population," *Nature*, vol. 464, pp. 773–777, apr 2010.

[71] M. Nicolae, S. Mangul, I. I. Măndoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from RNA-Seq data," *Algorithms for Molecular Biology*, vol. 6, p. 9, dec 2011.

[72] B. Li and C. N. Dewey, "RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, p. 323, dec 2011.

[73] H. Jiang and W. H. Wong, "Statistical inferences for isoform expression in RNA-Seq.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1026–32, apr 2009.

[74] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nature Biotechnology*, vol. 33, pp. 290–295, mar 2015.

[75] N. Nariai, K. Kojima, T. Mimori, Y. Sato, Y. Kawai, Y. Yamaguchi-Kabata, and M. Nagasaki, "TIGAR2 : sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads," *BMC Genomics*, vol. 15, p. S5, dec 2014.

[76] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology*, vol. 34, pp. 525–527, may 2016.

[77] R. Patro, S. M. Mount, and C. Kingsford, "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms," *Nature Biotechnology*, vol. 32, pp. 462–464, may 2014.

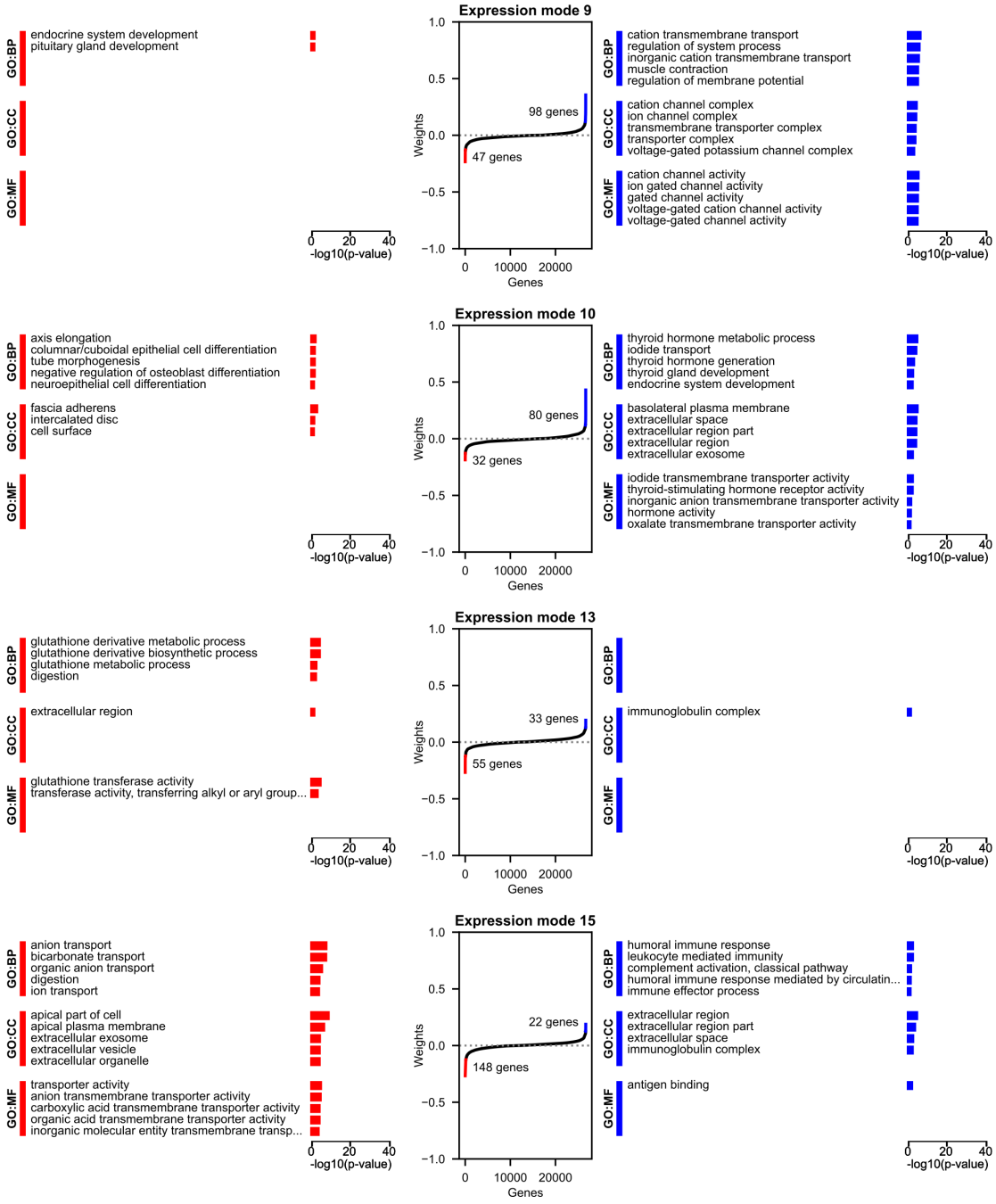
[78] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression.," *Nature methods*, vol. 14, pp. 417–419, apr 2017.

[79] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek, "GENCODE reference annotation for the human and mouse genomes," *Nucleic Acids Research*, vol. 47, pp. D766–D773, jan 2019.

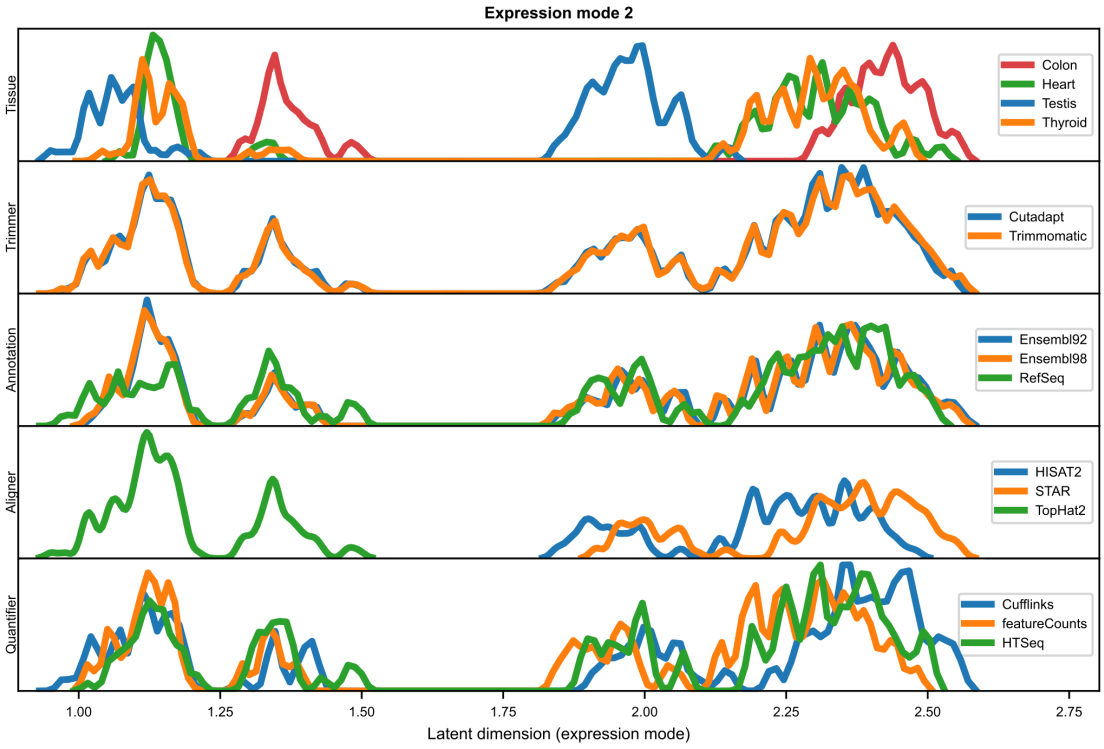
3.8 Supplementary figures and tables

Tissue	Sample ID
Colon	ERR315348
Colon	ERR315357
Colon	ERR315400
Colon	ERR315462
Heart	ERR315328
Heart	ERR315331
Heart	ERR315356
Heart	ERR315384
Testis	ERR315351
Testis	ERR315352
Testis	ERR315456
Testis	ERR315492
Thyroid	ERR315397
Thyroid	ERR315412
Thyroid	ERR315422
Thyroid	ERR315428

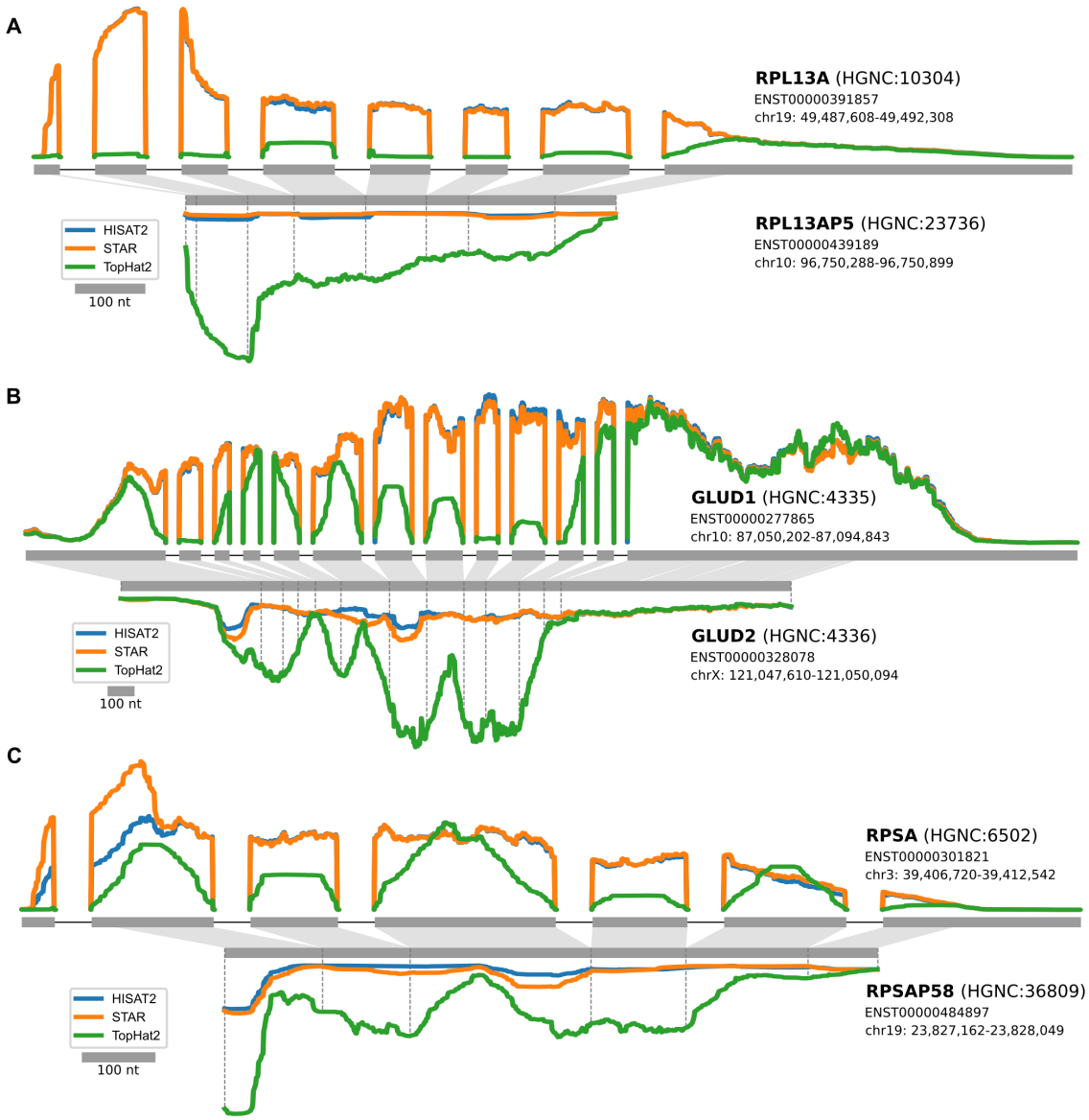
Table 1 – Samples tissues and ID from the Array-Express E-MTAB-2836 datasets (https ://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2836/)



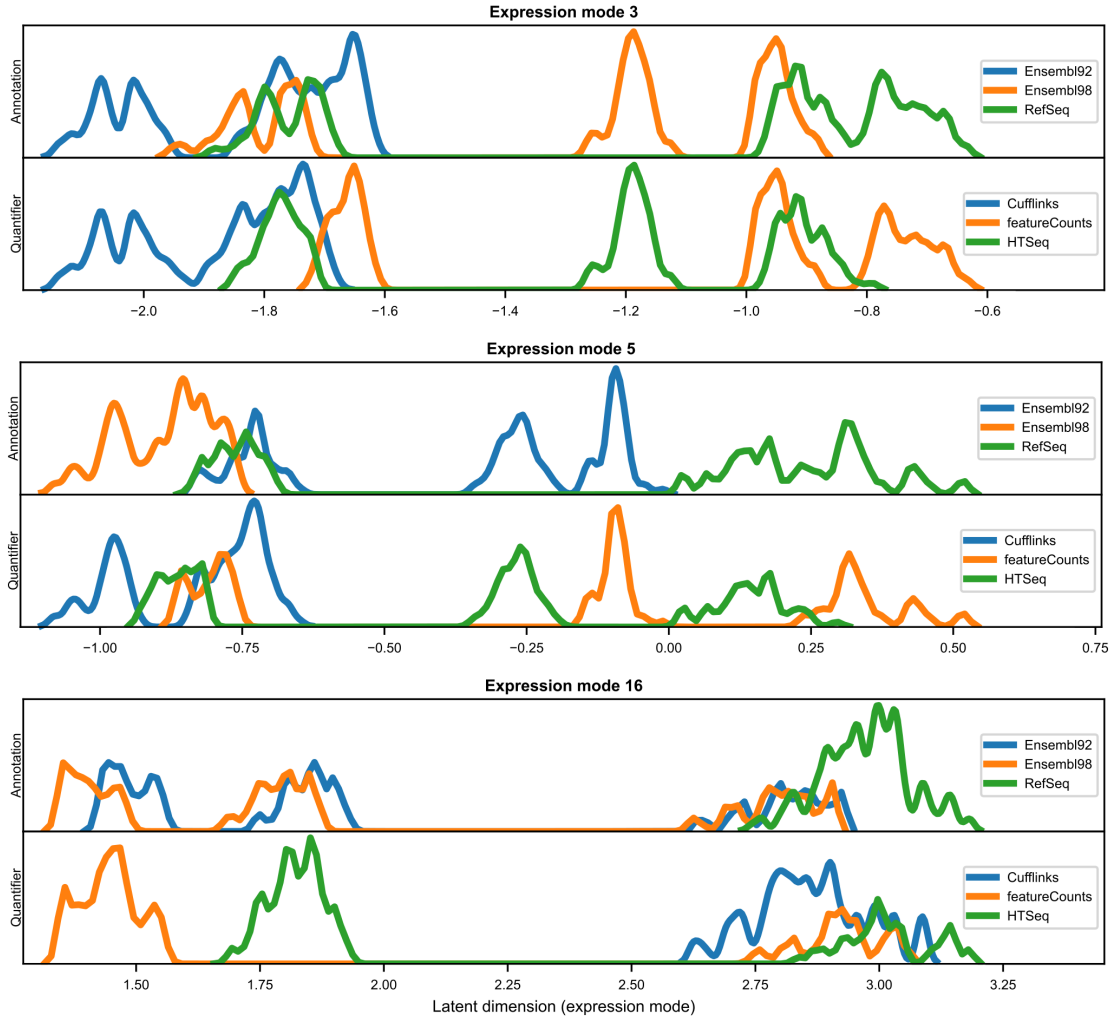
Supplementary Figure 1 – The gene weight distributions and the two-sided Gene Ontology (GO) enrichment analysis for the four biological modes presented in Figure 2C are displayed here. For every biological mode, the side with the largest number of genes also has GO terms that are related to the tissue that the biological mode is able to classify.



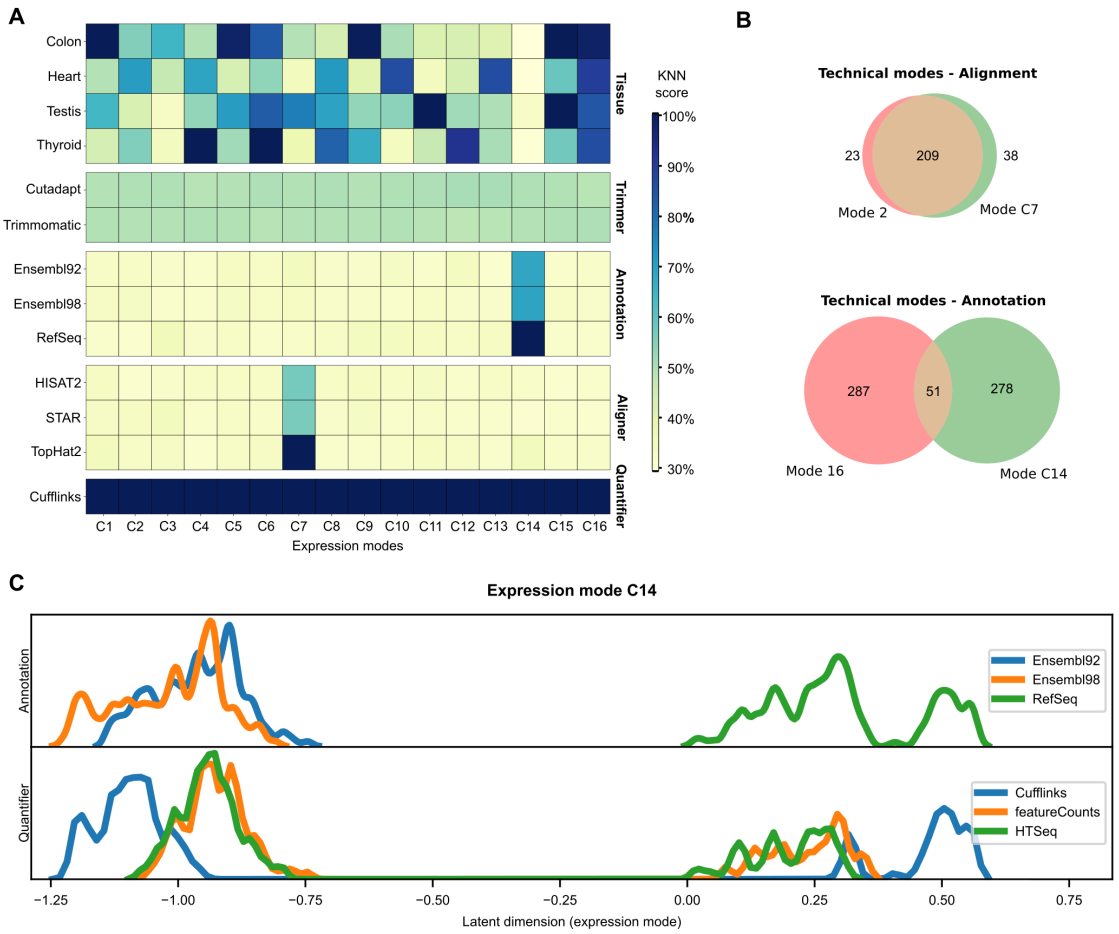
Supplementary Figure 2 – Distribution of the pipelines along the latent dimension of the expression mode 2 (EM2). The same distribution is shown in five occurrences, one for each pipeline step. Within each occurrence, the pipelines are separated according to the tool used. This representation helps us interpret the information used by the expression mode. For EM2, we can observe that the distribution is mainly driven by the alignment software, since TopHat2 is clustered alone, clearly separated from STAR and HISAT2. We can also see that some of the variability in the shape of the alignment clusters is due to biological variability between tissues, with clusters of colon and testis tissues clustering apart from the other tissues. We might interpret this as if some important genes in defining the alignment clustering have large expression differences in these tissues.



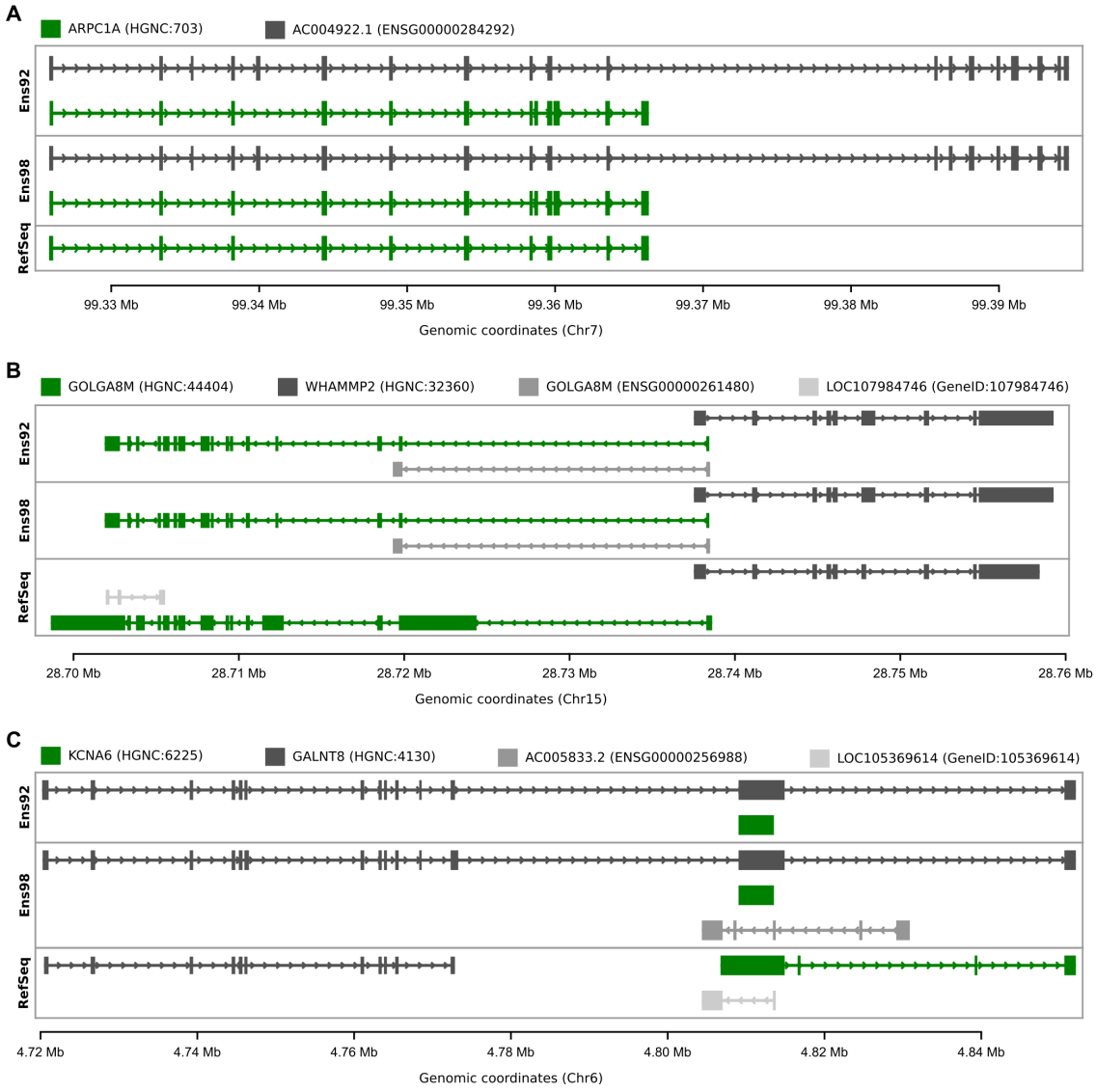
Supplementary Figure 3 – Comparison of the read profiles along paired genes and pseudogenes for TopHat2, HISAT2 and STAR. This representation is the same as described in Figure 3F.



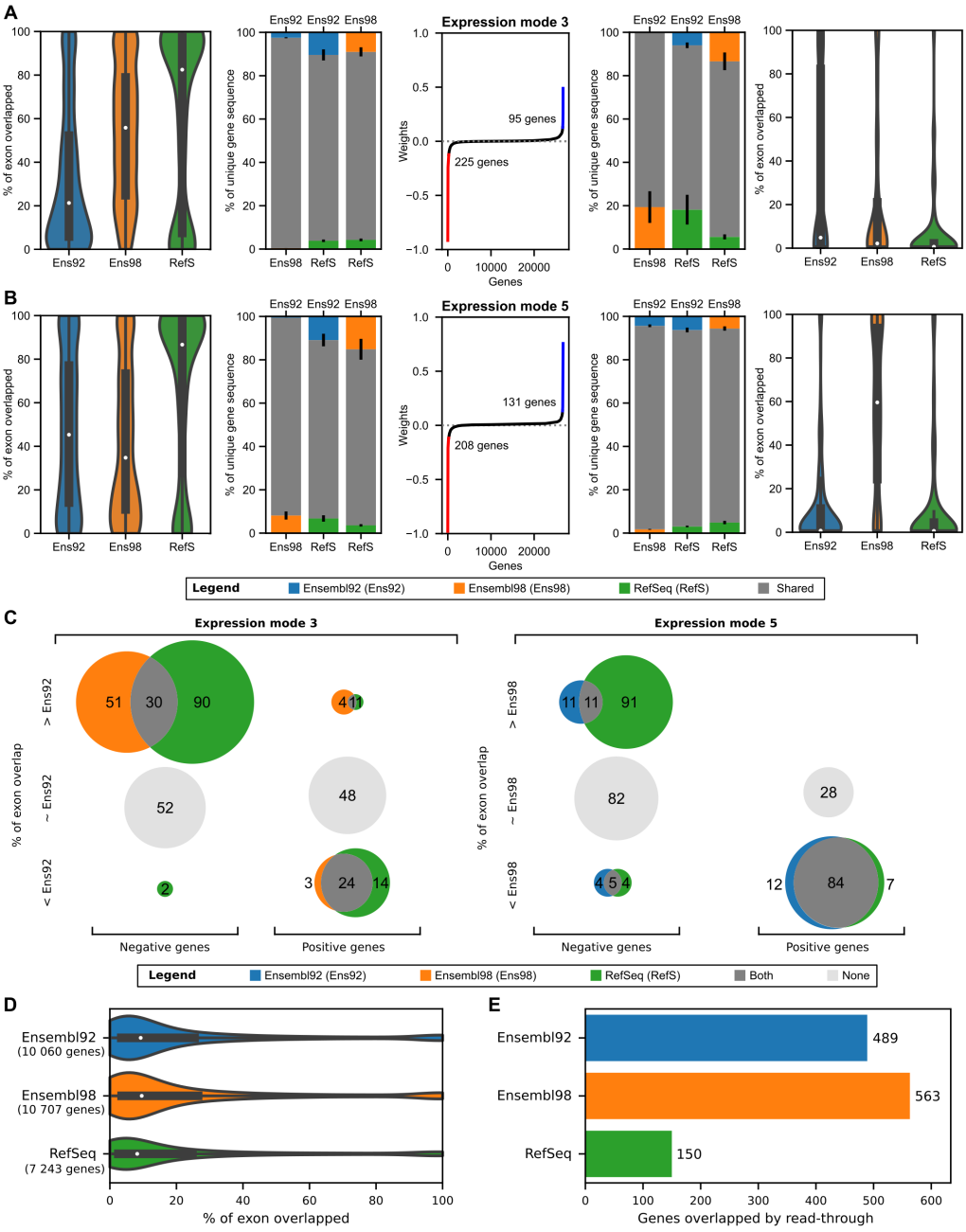
Supplementary Figure 4 – Expression modes 3, 5 and 16 are all partly clustered for the genome annotation and the quantifier tools. Distributions of the pipeline along the expression modes are presented for the annotation and the quantifier. In every distribution, Cufflinks and a genome annotation are clustered on one side, whereas the two other quantifiers and annotations are separated between the two sides.



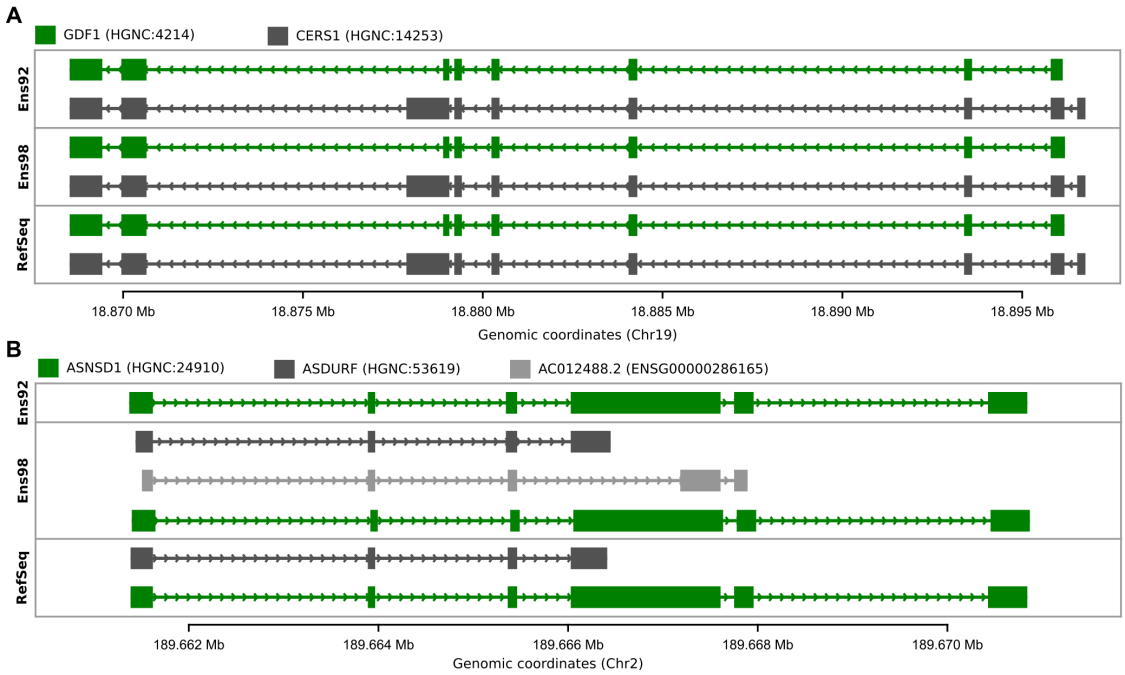
Supplementary Figure 5 – Results and observations of an ICA model that was computed by exclusively using expression datasets that were quantified through Cufflinks. **A** presents the KNN score heatmap for expression mode classification. Only two technical modes, C7 and C14 were detected, respectively linked to aligners and genome annotations. **B** quantifies the overlap of significant genes between technical modes linked to the alignment and annotation from this model to those from the original ICA model. **C** illustrates the distribution of pipelines along C14. In this instance, expression datasets from featureCounts and HTSeq were projected along the EMC14, even if they were not part of the model datasets.



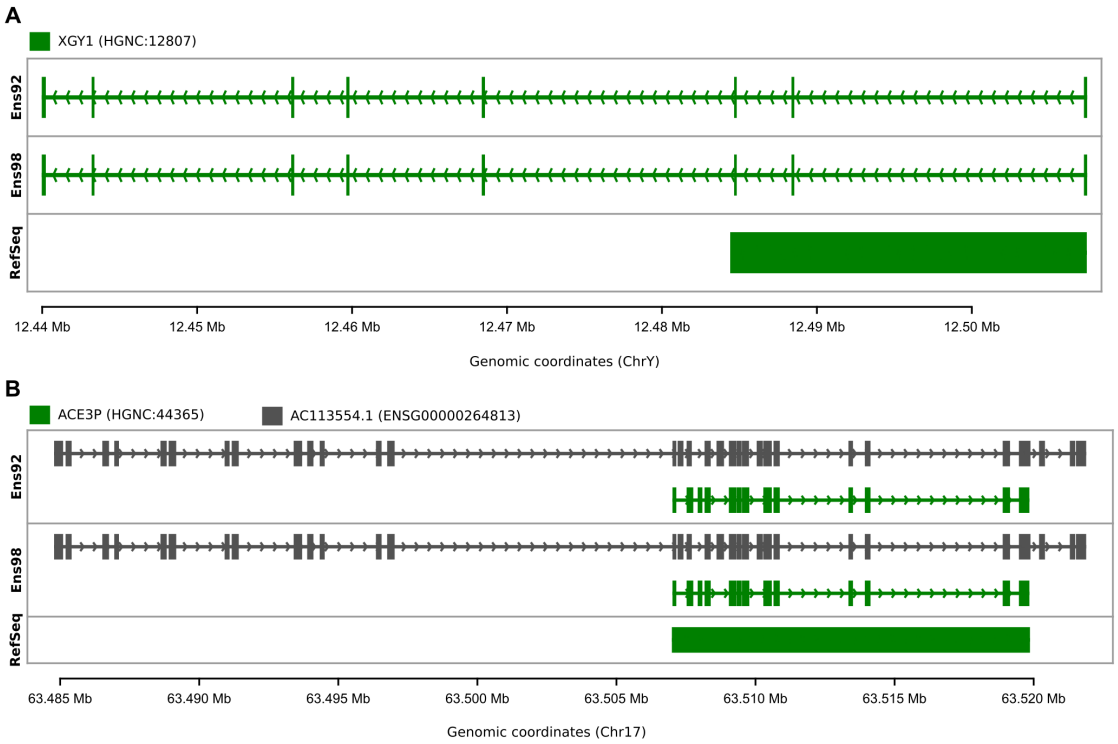
Supplementary Figure 6 – Each gene is summarized as a single entry with a single level of information, where the displayed structure of the gene is a one-dimensional projection, the shadow, of all the different isoforms. Each position in a block is part of an exon in a least one isoform, whereas each position not in block is an intron in all the different isoforms. The plots are centered on a gene of interest, the green gene, and display all the other genes that share at least one chromosomal coordinate in at least one annotation. Each gene of interest is displayed in the three studied genome annotations, for comparison. The genes are identified by their symbol, and the most common identifier (e.g. a HGNC identifier if they possess one, or their genome annotation specific identifier in the second case). The three genes of interest were found in a technical mode related to the classification between Ensembl and RefSeq. ARPCA1 (**A**) is a protein-coding gene found in EM16, GLOGA8M (**B**) is a protein-coding gene found in EMC14 and KCNA6 (**C**) is a protein-coding gene found in both EM16 and EMC14.



Supplementary Figure 7 – **A** and **B** present the plots described in Figure 4 for EM3 and EM5, highlighting the features discriminating the different genome annotations. **C** classifies the positive and negative genes from EM3 and EM5 according to their percentage of exon overlapped, relative to the genome annotation that is clustered individually. Genes with a score within 10% of the reference annotation score, or with two bigger but opposite scores, were identified as being similar to the reference annotation. Genes with score at least 10% bigger or smaller than the reference annotation were classified as such, colored accordingly to which scores were significantly different. **D** quantifies the percentage of exon overlapped by other genes, either on the sense or antisense strands, for all the 26 713 genes considered in this study. Only the genes with non-null overlap were used to produce the violin plots, and their quantity is presented with the genome annotation names. **E** quantifies the number of read-through genes overlapping at least one gene included in this study.



Supplementary Figure 8 – **A.** CERS1 and GDF1 are two overlapping genes that share a large proportion of their exons, and that are annotated in the same manner across the three studied annotations. **B.** ASNSD1 transcripts were split into three different genes from Ensembl version 92 to version 98. One of the new genes also has an HGNC ID, and is found in RefSeq. The gene representation is the same as described in Supplementary Figure 5.



Supplementary Figure 9 – A and B represent pseudogenes that are differently annotated in RefSeq and Ensembl. The gene representation is the same as described in Supplementary Figure 5.

Chapitre 4
Discussion

L'étude méthodologique amène nécessairement son lot de questions sur l'application optimale d'une technique. Ce mémoire n'a pas pour objectif de se prononcer sur l'annotation génomique permettant d'obtenir les meilleurs résultats, ni sur le pipeline bio-informatique de séquençage de l'ARN optimal. L'objectif de ce mémoire est plutôt la reconnaissance de l'importance du choix d'annotation génomique en bio-informatique, en étudiant plus précisément son impact sur la quantification des gènes en RNA-seq. Avec la notion d'importance du choix, il y a aussi celle de la non-trivialité de ce dernier. Les projets d'annotation actuels s'opposent par des hypothèses biologiques implicites différentes, une diversité d'outils et de logiciels, de méthode de travail, d'intégration de preuves biologiques. Les outils qui nous permettraient de déterminer qu'une annotation serait potentiellement plus intéressante à utiliser qu'une autre nous permettraient aussi de corriger l'annotation, et donc de retourner au point de départ sur la question de la sélection, admettant un certain décalage temporel. Peu importe le niveau de convergence des annotations, il restera toujours une différence identifiable tant que non complètement identiques, mais l'important est l'ordre de grandeur de cette différence.

4.1 Les annotations génomiques ont un impact dans le RNA-seq

Les deux articles présentés dans ce mémoire permettent de mieux définir la réalité des annotations génomiques dans la méthodologie de RNA-seq. Le premier article met en lumière que la communauté n'estime pas les annotations génomiques comme étant aussi importantes méthodologiquement que les logiciels d'alignement et de quantification. Si les annotations étaient considérées comme importantes, celles-ci seraient méthodologiquement décrites. Seulement 50% des articles ont précisé l'origine de l'annotation génomique utilisée, contre 88% et 74% pour les identifications de logiciels d'alignement et de quantification. Le deuxième article démontre qu'il est possible de classifier les résultats de quantification de RNA-seq en fonction de l'annotation utilisée, même entre deux versions de la même annotation génomique. Pour qu'il y ait classification, il doit y avoir un biais de quantification stable et retrouvé à travers les différents tissues et outils inclus dans l'étude. L'article démontre aussi que les annotations affectent la quantification d'un nombre de

gènes similaire aux outils de quantification, faisant ainsi de ces deux étapes celles qui ont le plus large impact, pour les outils testés.

L'important ici est l'accumulation des conclusions des deux articles. S'il s'était avéré que les annotations n'avaient aucun impact important sur la quantification, il ne serait pas aussi important de rapporter l'information méthodologique. Mais la problématique méthodologique est aggravée puisqu'on démontre qu'il y a un biais important de quantification, et que celui-ci est majoritairement ignoré dans la communauté.

L'impact des annotations sur la quantification identifié dans le second article a été classifié comme étant intrinsèque dans le cas où la définition même du gène est changeante, et donc influence les lectures comptabilisées, ou extrinsèque dans le cas où la définition d'autres gènes environnants interagissait avec la quantification du gène d'intérêt. Cette classification met en lumière une diversité, certes binaire, de différences identifiables. Ces deux catérogies ajoutent des biais de quantification différents, où les facteurs extrinsèques semblent pouvoir être corrigés par certaines approches de quantification, alors que les facteurs intrinsèques semblent totalement indépendants.

4.2 Les annotations génomiques résument nos connaissances biologiques

Les annotations génomiques sont le lieu de rassemblement de nos connaissances sur la biologie des gènes à l'échelle du génome et illustrant donc certains mécanismes de biogenèse des ARN, soit la transcription et l'épissage. Lincoln Stein (2001) a décrit l'acte d'annoter les génomes avec un parallèle de l'étude de la Torah. Le Talmud, issu du travail d'interprétation de la Torah, tente d'en distiller la substance, l'intention, en y construisant un système de loi. Stein décrit la Torah comme un ensemble cryptique, contradictoire et irrégulier. Aussi, du Talmud ressort l'adjectif talmudique, qui décrit par extension quelque chose caractérisée par des détails fins et subtils.

Bien que le parallèle entre la Torah et le génome, dont les adjectifs de cryptique, contradictoire et irrégulier ont une valeur tant que notre compréhension de ces entités n'est pas parfaite, offre une certaine compréhension de la problématique d'interprétation du génome, l'analogie perd son sens à l'annotation. Alors que l'annotation décrit les éléments composants du génome, le Talmud se veut décrire lois et traditions issues de l'interprétation

de la Torah. On est donc à un degré supérieur, le degré qu'il manque présentement aux annotations génomiques. Bien qu'il soit d'importance de correctement identifier les différents produits de transcription possibles, il est encore plus important de les relier sur un niveau supérieur à l'intention explicite d'annotation derrière ceux-ci. Les projets d'annotation sont des projets scientifiques, mais ils possèdent un lot d'opacité et de décisions implicites qui rendent difficile la critique, l'interprétation et l'amélioration potentielle de ces projets. La section prochaine détaille certains exemples reliés à ce manque informationnel, en mettant l'accent sur des exemples provenant du chapitre 3 et d'observations autres.

4.2.1 La normalité du génome

Une question essentielle à tout projet d'annotation est de définir ce qui doit être annoté, soit la nature même du projet. Aken et al. (2016) décrivent le but du projet d'annotation d'Ensembl comme étant l'identification des éléments fonctionnels, en donnant en exemple la transcription d'ARNm, et des divers facteurs de régulation et d'expression. Considérant que ce qui est annoté a un impact sur soi-même et sur potentiellement d'autres gènes, il est important de se questionner sur les limites de ce que l'on considère comme étant fonctionnel, et donc annotable.

Une hypothèse du développement du cancer est l'accumulation progressive de diverses capacités qui lui permettent de déréguler sa prolifération cellulaire, de résister à la mort cellulaire, d'envahir son milieu agressivement (Hanahan & Weinberg, 2011). Cette dérégulation, ou plutôt changement d'équilibre de régulation, prend moléculairement action d'une grande diversité d'origine, passant par une accumulation de mutations, des réarrangements chromosomiques, des modifications d'épissage et autres (Hanahan & Weinberg, 2011). Les réarrangements chromosomiques, essentiellement modifications de la séquence de chromosomes, peuvent prendre une grande variété de formes, passant par la délétion, l'ajout ou l'inversion de séquences, jusqu'à la translocation de sections complètes de chromosomes. Ces divers réarrangements peuvent affecter l'expression de divers ARN, dérégulant ainsi des processus biologiques (Zhang et al. 2018). Ces réarrangements peuvent créer des transcrits aberrants, qui ne sont pas retrouvés à l'extérieur de ce contexte de cancer et dont la séquence ne se retrouve pas de façon contiguë, admettant l'épissage d'introns, dans le génome de référence (Kowalski et al., 1999). Ces transcrits aberrants ne seront donc pas annotés, car on ne peut les positionner directement dans le génome, dans le contexte du format d'annotation actuel. Mais les réarrangements chromosomiques peuvent aussi

affecter les éléments cis-régulateurs d'un gène, pouvant être la cause d'une expression anormale de transcrits, mais qui peuvent potentiellement être repositionnés dans le génome, et donc annotés (Harewood & Fraser, 2014). Plusieurs cancers sont aussi caractérisés par une accumulation de mutations dans des séquences de facteurs d'épissage, ce qui produit des événements d'épissage aberrants, et donc des transcrits aberrants (Seiler et al., 2018). Puisque cette problématique voit le jour post-transcriptionnellement, et qu'elle n'est pas due à une modification des gènes parents des transcrits aberrants, il serait facile de les identifier comme étant des transcrits normaux des gènes en question.

L'interrogation soulevée ici est la question d'annotation de la normalité, du statut d'être sain. Avec l'hypothèse d'accumulation de caractéristiques du cancer, il est bien possible d'avoir un échantillon cellulaire caractérisé comme sain, mais qui possède moléculairement quelques propriétés du cancer. Ainsi, un tissu sain pourrait avoir des facteurs d'épissage mutés, soit assez pour produire des transcrits aberrants, mais pas assez pour être classifié comme non sain. Qui plus est, si le désir d'annotation est d'inclure tous les transcrits possibles, peu importe le statut mutationnel et l'intégrité des chromosomes, on se trouverait devant deux problématiques importantes. Premièrement, l'acceptation du non sain veut dire acceptation théorique de toutes différences fonctionnelles dans les différents acteurs de la biogenèse de ARN, ce qui veut dire que tout transcrit existe avec son lot potentiel de transcrits, probablement anecdotiques, divergents. Ceci complique la quantification des transcrits, et va à l'encontre de l'hypothèse que la plupart des gènes n'ont qu'un transcrit dominant (Ezkurdia et al., 2015). Deuxièmement, le modèle de données actuel définissant les annotations génomiques n'est pas compatible avec les mécanismes de divergence de séquence du génome de référence. Pour inclure l'étude systématique des réarrangements chromosomiques annotés, il faudrait pouvoir les inclure dans les annotations et outils.

Mais le point important n'est pas tant la décision prise, que de prendre explicitement la décision concernant ce qui devrait être annoté. Ne pas prendre cette décision c'est aussi ne pas en évaluer les conséquences. Nous soulignons le besoin de références plus explicites dans leurs hypothèses biologiques. Les prochaines sections illustreront plus directement diverses problématiques observées en lien avec des décisions implicites d'annotation.

4.2.2 Les gènes de fusion

Les gènes de fusion, entre autres discutés dans le chapitre 3, sont aussi en lien avec la section précédente concernant le statut de normalité. Les gènes de fusion peuvent être classifiés en deux groupes, soit les trans et les cis. Les gènes de fusions trans combinent la séquence de deux gènes qui ne sont pas normalement positionnés de manière adjacente dans le génome, donc issus d'un réarrangement chromosomique. Les gènes de fusion cis sont eux génomiquement adjacents (Zhang et al., 2012). Les gènes de fusions sont des événements souvent associés au cancer, et une récente étude démontre que les gènes de fusion ne sont majoritairement qu'anecdotiquement détectés dans les tissus sains (Babiceanu et al., 2016). La majorité des gènes de fusion ne sont détectés que dans un seul échantillon parmi les 300 échantillons considérés dans l'étude, et l'article démontre aussi la présence de gènes de fusion issus de la translocation de chromosomes, ce qui remet définitivement en doute le statut sain des échantillons. Nous avons illustré qu'Ensembl possède un nombre de gènes de fusion cis annoté beaucoup plus élevé que RefSeq, et que ce nombre est en augmentation (Simoneau et al., 2020). Les gènes de fusion cis produisent des biais de quantification extrinsèque très importants, puisqu'ils occupent les mêmes coordonnées chromosomiques que deux autres gènes, ce qui rend la distribution des lectures plus difficile, surtout dans le cas de logiciel de quantification basé sur la distribution de compte, sans modèle d'optimisation (Simoneau et al., 2020). Comme illustré dans la méta-analyse méthodologique, au moins 25% des articles publiés, considérant HTSeq et featureCounts, se basent sur l'utilisation de tel logiciel de quantification, ce qui crée un biais pour tous les gènes possédant un gène de fusion cis (Simoneau et al., 2019).

Les questionnements soulevés sont la raison de divergence en termes de nombre de gènes de fusion dans chaque annotation, où Ensembl en possède environ trois fois plus que RefSeq, et l'avantage d'avoir ces annotations dans le contexte de leur biais potentiel. Considérant la nature à haut débit du RNA-seq, il est fort difficile d'évaluer l'importance de ce qu'on ne voit pas, puisque la large quantité de données ne permet que très peu l'exploration de celles-ci. Ainsi il est difficile de quantifier l'impact de la présence ou de l'absence de ces annotations dans la littérature sans réanalyser les diverses expériences. D'un autre côté, la divergence d'annotation devrait pouvoir s'expliquer, malgré le fait qu'on ne possède pas cette information. Cela pourrait être expliqué par l'utilisation de données ou logiciels différents, avec des hypothèses biologiques divergentes. L'important reste que les personnes utilisatrices soient au courant de la nature des données utilisées, et de l'impact de celles-ci. Considérant l'expression anecdotique de la grande majorité des gènes de fusion, pourrait-on dire que l'acte de les annoter apporte simplement du biais

pour les gènes chevauchés ?

4.2.3 Des gènes identiques

En quantifiant le chevauchement des gènes à l'intérieur d'une même annotation, une situation bien curieuse s'est présentée à nous, soit l'existence de gènes parfaitement identiques chez Ensembl version 98. Non seulement ces paires de gènes possèdent la même séquence, mais ils partagent aussi exactement les mêmes coordonnées chromosomiques. Résumés dans le tableau 4.1, ces neuf pairs de gènes sont tous des pseudogènes de l'ARN ribosomal RNA5S. Les gènes identifiés comme nouveaux gènes n'étaient pas présents dans Ensembl version 92, mais le sont dans Ensembl version 98. Étrangement, la plus vieille copie de ces gènes possède le biotype d'ARN ribosomaux, alors que la plus récente possède le biotype de pseudogène d'ARN ribosomaux, bien que le symbole des gènes utilise la nomenclature des pseudogènes, soit le symbole du gène parent, suivi d'un P et d'un nombre.

Symbole	Gène Ensembl	Nouveau gène	Coordonnées
RNA5SP71	ENSG00000199837	ENSG00000285609	1 : 182,944,365-182,944,490
RNA5SP85	ENSG00000202248	ENSG00000285626	2 : 11,561,661-11,561,779
RNA5SP150	ENSG00000199839	ENSG00000285546	3 : 181,822,872-181,822,990
RNA5SP172	ENSG00000201931	ENSG00000285574	4 : 177,457,111-177,457,228
RNA5SP194	ENSG00000201532	ENSG00000285956	5 : 139,012,329-139,012,441
RNA5SP250	ENSG00000199404	ENSG00000231139	7 : 152,592,973-152,593,091
RNA5SP443	ENSG00000199953	ENSG00000285757	17 : 45,327,366-45,327,497
RNA5SP463	ENSG00000199900	ENSG00000285674	19 : 7,886,977-7,887,096
RNA5SP506	ENSG00000272351	ENSG00000285776	X : 69,672,479-69,672,597

Tableau 4.1 – Pseudogènes d'ARNr dupliqués. Liste de pseudogènes d'ARN ribosomal possédant une copie identique du gène identifié sous un autre nom. Les gènes notés comme nouveaux ont été rajoutés dans l'annotation d'Ensembl entre les versions 92 et 98.

Encore une fois, il s'agit de la question d'intention. Y a-t-il le désir volontaire d'avoir ces gènes identiques dans la même annotation ? Ou est-ce une erreur de pipeline d'annotation qui s'est introduite ?

4.3 Quelles sont les décisions à prendre ?

Le seul fait clair et indiscutable que cette section offre est la nécessité de rapporter intégralement l'ensemble des décisions prises en rapport au choix de référence et à la potentielle

modification de celle-ci. Outre ce besoin d'exactitude méthodologique qui permet la réplication et l'exploration indépendante des résultats, on ne peut émettre de recommandations *sine qua non*, mais illustrer quelques choix possibles.

4.3.1 L'annotation et sa version

Le premier choix évident est la source de l'annotation génomique et son numéro de version. Bien qu'il soit difficile d'indiquer l'annotation génomique offrant les meilleurs résultats, il est possible de faire un choix en lien avec la facilité d'utilisation. La méta-analyse méthodologique a démontré que les personnes utilisatrices d'Ensembl et de GENCODE avaient beaucoup plus tendance à indiquer la version de leur annotation comparativement à RefSeq, l'hypothèse avancée étant la nature de la présentation de l'information sur les ressources (Simoneau et al., 2019). Les pages individuelles d'Ensembl offrent aussi une présentation plus moderne et plus interactive des gènes, permettant ainsi d'explorer certaines questions en rapport aux résultats d'annotation. Bien que les informations d'annotation sont difficiles à comparer dans un contexte biologique complexe, le choix d'annotation à utiliser devrait aussi compter le reste des ressources autour de l'information.

Les annotations génomiques, ainsi que les gènes les composant, sont versionnées. Nous mettons en garde contre l'utilisation du numéro de version des gènes individuels, puisqu'un gène pourrait conserver la même définition, et donc le même numéro de version, à travers plusieurs versions d'annotation, mais changer de quantification en fonction des gènes environnants qui eux se sont vus modifier, dû aux biais extrinsèques de la quantification. Ainsi il ne suffit pas de regarder le gène d'intérêt, mais bien de le conserver dans son contexte.

4.3.2 Indice de qualité des annotations

Les annotations génomiques offrent divers indices de qualité concernant les objets annotés, soit le système TSL pour Ensembl, et la séparation prédiction et curation manuelle pour RefSeq. Bien qu'il soit théoriquement possible de n'utiliser qu'un sous-ensemble, soit les données les plus fiables, d'une annotation, ce n'est pas quelque chose qui a été observé dans la méta-analyse méthodologique. Qui plus est, aucune référence d'annotation ne distribue de fichier contenant qu'une partie filtrée de l'ensemble. À notre connaissance, il n'existe pas non plus d'étude s'étant intéressée à cette question.

4.3.3 Quelques recommandations

Éviter les ressources clef en main. La méthodologie en RNA-seq n'est pas rendue à un point où l'on peut simplement effectuer ses calculs à partir d'un pipeline unique, sans le questionner. Les ressources clef en main peuvent aussi inclure des étapes ou données difficiles à détailler pour la personne utilisatrice dans sa méthodologie. De même pour les logiciels commerciaux qui peuvent reposer sur du code inaccessible.

Éviter les sources alternatives de données. Il est possible de retrouver les diverses annotations génomiques sur d'autres sites internet ou serveurs que ceux les publiant originellement. Il est fortement déconseillé d'utiliser ces sources alternatives, car l'information s'y retrouvant n'est peut-être pas intégrale, ni à jour.

Comparer les résultats. Une façon, certes chronophage, d'aborder une multiplication des ressources sans aide à la décision est l'utilisation parallèle de ces diverses ressources, avec une comparaison finale des résultats. Par exemple, une étude d'expression différentielle pourrait être effectuée avec Ensembl et RefSeq indépendamment, puis les groupes de gènes significatifs pourraient être comparés. Ainsi cela donne deux groupes de gènes restreints à explorer manuellement, en regardant ce qui est similaire et divergent entre les deux groupes, donnant ainsi une meilleure compréhension des résultats et de leurs limites.

Et il ne faut pas oublier que le plus grand piège de la bio-informatique, c'est de ne pas comprendre ce que l'on fait. Il est facile de faire rouler un logiciel, d'utiliser une ressource ou une autre, mais si l'on ne connait pas les différences entre celles-ci, si l'on ne connait pas les différentes hypothèses sous-jacentes, nos résultats ont bien peu de valeur.

Conclusions

Max Delbrück, biophysicien germano-américain, énonce le principe du *Limited Sloppiness*, soit de la négligence limitée, pour décrire son approche de la science (Delbrück, 1978). Ce principe illustre qu'une démarche trop rigoureuse et rigide laisse peu de place à l'exploration de l'imprévisible. On peut aussi décrire la biologie comme un système négligeant, où les erreurs d'une machinerie moléculaire imparfaite créent une diversification essentielle pour répondre à des problématiques encore potentiellement inconnues, et donc de s'adapter, d'évoluer. Mais il est important que cette négligence soit limitée, soit d'un ordre de grandeur permettant de conserver l'intégrité du système biologique, ou permettant une reproductibilité des résultats en science. Par conséquent, ces systèmes se doivent d'être robustes, mais aussi souples.

Ce mémoire démontre l'importance du choix d'annotation génomique en séquençage de l'ARN, illustrant que ce choix est peu explicitement rapporté, mais qu'il est responsable d'une large part des biais des quantifications. Ce mémoire discute aussi de l'impossibilité de la tâche d'annotation. Comme la biologie est un système partiellement négligeant, peut-on s'attendre à en expliquer l'entièreté par des modèles et règles simples ? Les projets d'annotation tels que décrits présentement sont des approches réductionnistes à la question d'identification des éléments fonctionnels du génome. Outre les problématiques de l'approche réductionniste en rapport au concept d'émergence en biologie (Longo et al., 2012), il est impossible de fixer un système dans des règles, lorsque ce dernier n'est pas stable. Alors que la physique met de l'avant l'universalité de ses lois élémentaires, la biologie moléculaire est contextualisée.

Bien que les annotations génomiques occupent un rôle crucial en tant qu'outil dans un large nombre d'analyses, leur rôle d'accumulation du savoir est limité par son modèle restrictif d'outil. La réponse à plusieurs limitations des annotations pourrait se trouver dans un abandon du modèle réductionniste, en faveur d'un modèle plutôt probabiliste et complexe, où l'interaction des mécanismes serait mise de l'avant, plutôt que l'étude directe de leurs résultats, soit les transcrits.

Mais en attendant, la méthodologie des analyses effectuées peut simplement être améliorée en faisant des choix informés, et en les rapportant convenablement.

Références

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., Vogel, J. H., White, S., Zadissa, A., Flicek, P., et Searle, S. M. (2016) The Ensembl gene annotation system. *Database : the journal of biological databases and curation*, 2016 : 1–19.

Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y., Lazar, I. M., et Li, H. (2016) Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Research*, 44(6) : 2859–2872.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., et Yeh, L. S. L. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(DATABASE ISS.) : 154–159.

Ballouz, S., Dobin, A., et Gillis, J. A. (2019) Is it time to change the reference genome ? *Genome Biology*, 20(1) : 1–9.

Ballouz, S., Dobin, A., Gingeras, T. R., et Gillis, J. (2018) The fractured landscape of RNA-seq alignment : the default in our STARs. *Nucleic acids research*, 46(10) : 5125–5138.

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T. P., Reecy, J. M., et Tuggle, C. K. (2019) Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genomics*, 20(1) : 1–19.

Boguski, M. S., Lowe, T. M., et Tolstoshev, C. M. (1993) dbEST - database for "expressed sequence tags". *Nature Genetics*, 4(august) : 332–333.

Boivin, V., Deschamps-Francoeur, G., Couture, S., Nottingham, R. M., Bouchard-Bourelle, P., Lambowitz, A. M., Scott, M. S., et Abou-Elela, S. (2018) Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *RNA*, 24(7) : 950–965.

Brenner, S., Jacob, F., et Meselson, M. (1961) An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature*, 190 : 576.

Brent, M. (2007) GTF2.2 : A Gene Annotation Format. `http://mblab.wustl.edu/GTF22.html`.

Brunet, M. A., Brunelle, M., Lucier, J. F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J. D., Dufour, P., Jacques, J. F., Fournier, I., Ouangraoua, A., Scott, M. S., Boisvert, F. M., et Roucou, X. (2019) OpenProt : A more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Research*, 47(D1) : D403–D410.

Burge, C. et Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen. *Journal of Molecular Biology*, 268(1) : 78–94.

Camacho, M. P. (2019) The Central Dogma Is Empirically Inadequate . . .No Matter How We Slice It. *Philosophy, Theory, and Practice in Biology*, 11(6).

Cannata, N., Merelli, E., et Altmare, R. B. (2005) Time to organize the bioinformatics resourceome. *PLoS Computational Biology*, 1(7) : 0531–0533.

Chao, H. P., Chen, Y., Takata, Y., Tomida, M. W., Lin, K., Kirk, J. S., Simper, M. S., Mikulec, C. D., Rundhaug, J. E., Fischer, S. M., Chen, T., Tang, D. G., Lu, Y., et Shen, J. (2019) Systematic evaluation of RNA-Seq preparation protocol performance. *BMC Genomics*, 20(1) : 1–20.

Chen, R. et Butte, A. J. (2011) The reference human genome demonstrates high risk of type 1 diabetes and other disorders. *Pacific Symposium on Biocomputing 2011, PSB 2011*, pages 231–242.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H. C., Agarwala, R., McLaren, W. M., Ritchie, G. R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E. E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P., et Hubbard, T. (2011) Modernizing reference genome assemblies. *PLoS Biology*, 9(7) : 1–5.

Collins, F. S., Green, E. D., Guttmacher, A. E., et Guyer, M. S. (2003) A vision for the future of genomics research. *Nature*, 431(April) : 835–847.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et Mortazavi, A. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1) : 1–19.

Crick, F. (1970) Central Dogma of Molecular Biology. *Nature*, 227 : 561–563.

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., et Clamp, M. (2004) The Ensembl Automatic Gene Annotation System. *Genome Research*, (617) : 1–9.

Delbrück, M. (1978) Oral History Project : Max Delbruck, interviewed by Carolyn Harding, California Institute of Technology Archives. `http://oralhistories.library.caltech.edu/16/1/OH_Delbruck_M.pdf`.

Dupuis-Sandoval, F., Poirier, M., et Scott, M. S. (2015) The emerging landscape of small nucleolar RNAs in cell biology. *Wiley Interdisciplinary Reviews : RNA*, 6(4) : 381–397.

Ensembl (2020) Transcript Summary. `https://useast.ensembl.org/Help/View?id=151`.

Ezkurdia, I., Rodriguez, J. M., Carrillo-De Santa Pau, E., Vázquez, J., Valencia, A., et Tress, M. L. (2015) Most highly expressed protein-coding genes have a single dominant isoform. *Journal of Proteome Research*, 14(4) : 1880–1887.

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J., Kellis, M., Paten, B., Reymond, A., Tress, M. L., et Flicek, P. (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1) : D766–D773.

Frankish, A., Uszczynska, B., Ritchie, G. R., Gonzalez, J. M., Pervouchine, D., Petryszak, R., Mudge, J. M., Fonseca, N., Brazma, A., Guigo, R., et Harrow, J. (2015) Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*, 16(8) : 1–11.

Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., et Nilsson, P. (2009) Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics*, 10 : 1–14.

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason,

A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., Sverrisson, J. T., Gretarsdottir, S., Walters, G. B., Rafnar, T., Thjodleifsson, B., Bjornsson, E. S., Olafsson, S., Thorarinsdottir, H., Steingrimsdottir, T., Gudmundsdottir, T. S., Theodors, A., Jonasson, J. G., Sigurdsson, A., Bjornsdottir, G., Jonsson, J. J., Thorarensen, O., Ludvigsson, P., Gudbjartsson, H., Eyjolfsson, G. I., Sigurdardottir, O., Olafsson, I., Arnar, D. O., Magnusson, O. T., Kong, A., Masson, G., Thorsteinsdottir, U., Helgason, A., Sulem, P., et Stefansson, K. (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47(5) : 435–444.

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., et Shyr, Y. (2017) Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2) : 83–90.

Hanahan, D. et Weinberg, R. A. (2011) Hallmarks of cancer : The next generation. *Cell*, 144(5) : 646–674.

Harewood, L. et Fraser, P. (2014) The impact of chromosomal rearrangements on regulation of gene expression. *Human Molecular Genetics*, 23(R1) : 76–82.

Hogeweg, P. (2011) The roots of bioinformatics in theoretical biology. *PLoS Computational Biology*, 7(3) : 1–5.

Hood, L. et Galas, D. (2003) The digital code of DNA. *Nature*, 421(6921) : 444–448.

Hood, L. et Rowen, L. (2013) The human genome project : Big science transforms biology and medicine. *Genome Medicine*, 5(9).

Hopkin, K. (2009) The Evolving Definition of a Gene. *BioScience*, 59(11) : 928–931.

Hrdlickova, R., Toloue, M., et Tian, B. (2017) RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews : RNA*, 8(1).

Hsu, F., Kent, J. W., Clawson, H., Kuhn, R. M., Diekhans, M., et Haussler, D. (2006) The UCSC known genes. *Bioinformatics*, 22(9) : 1036–1046.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, 409(February).

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) : 931–945.

Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., Anthon, C., Beard, N., Berka, K., Bolser, D., Booth, T., Bretaudeau, A., Brezovsky, J., Casadio, R., Cesareni, G., Coppens, F., Cornell, M., Cuccuru, G., Davidsen, K., Vedova, G. D., Dogan, T., Doppelt-Azeroual, O., Emery, L., Gasteiger, E., Gatter, T., Goldberg, T., Grosjean, M., Grüning, B., Helmer-Citterich, M., Ienasescu, H., Ioannidis, V., Jespersen, M. C., Jimenez, R., Juty, N., Juvan, P., Koch, M., Laibe, C., Li, J.-W., Licata, L., Mareuil, F., Mičetić, I., Friborg, R. M., Moretti, S., Morris, C., Möller, S., Nenadic, A., Peterson, H., Profiti, G., Rice, P., Romano, P., Roncaglia, P., Saidi, R., Schafferhans, A., Schwämmle, V., Smith, C., Sperotto, M. M., Stockinger, H., Vařeková, R. S., Tosatto, S. C., de la Torre, V., Uva, P., Via, A., Yachdav, G., Zambelli, F., Vriend, G., Rost, B., Parkinson, H., Løngreen, P., et Brunak, S. (2016) Tools and data services registry : a community effort to document bioinformatics resources. *Nucleic Acids Research*, 44(D1) : D38–D47.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et Haussler, a. D. (2002) The Human Genome Browser at UCSC. *Genome Research*, 12(6) : 996–1006.

Kowalski, P. E., Freeman, J. D., et Mager, D. L. (1999) Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes. *Genomics*, 57(3) : 371–379.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., Van Sluys, M. A., Soltis, P. S., Xu, X., Yang, H., et Zhang, G. (2018) Earth BioGenome Project : Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, 115(17) : 4325–4333.

Longo, G., Montévil, M., et Kauffman, S. (2012) No entailing laws, but enablement in the evolution of the biosphere. *GECCO'12 - Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Companion*, pages 1379–1391.

Loveland, J. E., Gilbert, J. G., Griffiths, E., et Harrow, J. L. (2012) Community gene annotation in practice. *Database*, 2012 : 1–8.

Maglott, D. R., Katz, K. S., Sicotte, H., et Pruitt, K. D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Research*, 28(1) : 126–128.

Mitra, S. K. (1978) Recognition between codon and anticodon. *Trends in Biochemical Sciences*, 3(3) : 153–156.

Mouilleron, H., Delcourt, V., et Roucou, X. (2016) Death of a dogma : Eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Research*, 44(1) : 14–23.

Nachtergaele, S. et He, C. (2017) The emerging biology of RNA post-transcriptional modifications. *RNA Biology*, 14(2) : 156–163.

Nottingham, R. M., Wu, D. C., Qin, Y., Yao, J., Hunicke-Smith, S., et Lambowitz, A. M. (2016) RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *RNA*, 22(4) : 597–613.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., et Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI : current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1) : D733–D745.

Palazzo, A. F. et Lee, E. S. (2015) Non-coding RNA : What is functional and what is junk ? *Frontiers in Genetics*, 5(JAN) : 1–11.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T. J., Harrow, J., et Gerstein, M. B. (2012) The GENCODE pseudogene resource. *Genome Biology*, 13(9).

Pennisi, E. (2000) Ideas fly at gene-finding jamboree. *Science*, 287(5461) : 2182–2184.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y. C., Madugundu, A. K., Pandey, A., et Salzberg, S. L. (2018) CHESS : A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology*, 19(1) : 1–14.

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., Hart, E., Suner, M.-M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., DiCuccio, M., Kellis, M., Lee, J., Lin, M. F., Schuster, M., Shkeda, A., Amid, C., Brown, G., Dukhanina, O., Frankish, A., Hart, J., Maidak, B. L., Mudge, J., Murphy, M. R., Murphy, T., Rajan, J.,

Rajput, B., Riddick, L. D., Snow, C., Steward, C., Webb, D., Weber, J. A., Wilming, L., Wu, W., Birney, E., Haussler, D., Hubbard, T., Ostell, J., Durbin, R., et Lipman, D. (2009) The consensus coding sequence (CCDS) project : Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, 19 : 1316–1323.

Pyrkosz, A. B., Cheng, H., et Brown, C. T. (2013) RNA-Seq Mapping Errors When Using Incomplete Reference Transcriptomes of Vertebrates. *arXiv*.

Reyes, A. et Huber, W. (2018) Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*, 46(2) : 582–592.

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J. J., Lopez, G., Valencia, A., et Tress, M. L. (2013) APPRIS : Annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, 41(D1) : 110–117.

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., Bani Asadi, N., Gerstein, M. B., Wong, W. H., Snyder, M. P., Schadt, E., et Lam, H. Y. K. (2017) Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature communications*, 8(1) : 59.

Seiler, M., Peng, S., Agrawal, A. A., Palacino, J., Teng, T., Zhu, P., Smith, P. G., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, J. J., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Cho, J., DeFreitas, T., Frazer, S., Gehlenborg, N., Getz, G., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Meier, S., Noble, M. S., Saksena, G., Voet, D., Zhang, H., Bernard, B., Chambwe, N., Dhankani, V., Knijnenburg, T., Kramer, R., Leinonen, K., Liu, Y., Miller, M., Reynolds, S., Shmulevich, I., Thorsson, V., Zhang, W., Akbani, R., Broom, B. M., Hegde, A. M., Ju, Z., Kanchi, R. S., Korkut, A., Li, J., Liang, H., Ling, S., Liu, W., Lu, Y., Mills, G. B., Ng, K. S., Rao, A., Ryan, M., Wang, J., Weinstein, J. N., Zhang, J., Abeshouse, A., Armenia, J., Chakravarty, D., Chatila, W. K., de Bruijn, I., Gao, J., Gross, B. E., Heins, Z. J., Kundra, R., La, K., Ladanyi, M., Luna, A., Nissan, M. G., Ochoa, A., Phillips, S. M., Reznik, E., Sanchez-Vega, F., Sander, C., Schultz, N., Sheridan, R., Sumer, S. O., Sun, Y., Taylor, B. S., Wang, J., Zhang, H., Anur, P., Peto, M., Spellman, P., Benz, C., Stuart, J. M., Wong, C. K., Yau, C., Hayes, D. N., Parker, J. S., Wilkerson, M. D., Ally, A., Balasundaram, M., Bowlby, R., Brooks, D., Carlsen, R., Chuah, E., Dhalla, N., Holt, R., Jones, S. J., Kasaian, K., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Mungall, K., Robertson, A. G., Sadeghi, S., Schein, J. E., Sipahimalani,

P., Tam, A., Thiessen, N., Tse, K., Wong, T., Berger, A. C., Beroukhim, R., Cherniack, A. D., Cibulskis, C., Gabriel, S. B., Gao, G. F., Ha, G., Meyerson, M., Schumacher, S. E., Shih, J., Kucherlapati, M. H., Kucherlapati, R. S., Baylin, S., Cope, L., Danilova, L., Bootwalla, M. S., Lai, P. H., Maglinte, D. T., Van Den Berg, D. J., Weisenberger, D. J., Auman, J. T., Balu, S., Bodenheimer, T., Fan, C., Hoadley, K. A., Hoyle, A. P., Jefferys, S. R., Jones, C. D., Meng, S., Mieczkowski, P. A., Mose, L. E., Perou, A. H., Perou, C. M., Roach, J., Shi, Y., Simons, J. V., Skelly, T., Soloway, M. G., Tan, D., Veluvolu, U., Fan, H., Hinoue, T., Laird, P. W., Shen, H., Zhou, W., Bellair, M., Chang, K., Covington, K., Creighton, C. J., Dinh, H., Doddapaneni, H. V., Donehower, L. A., Drummond, J., Gibbs, R. A., Glenn, R., Hale, W., Han, Y., Hu, J., Korchina, V., Lee, S., Lewis, L., Li, W., Liu, X., Morgan, M., Morton, D., Muzny, D., Santibanez, J., Sheth, M., Shinbrot, E., Wang, L., Wang, M., Wheeler, D. A., Xi, L., Zhao, F., Hess, J., Appelbaum, E. L., Bailey, M., Cordes, M. G., Ding, L., Fronick, C. C., Fulton, L. A., Fulton, R. S., Kandoth, C., Mardis, E. R., McLellan, M. D., Miller, C. A., Schmidt, H. K., Wilson, R. K., Crain, D., Curley, E., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Thompson, E., Yena, P., Bowen, J., Gastier-Foster, J. M., Gerken, M., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Wise, L., Zmuda, E., Corcoran, N., Costello, T., Hovens, C., Carvalho, A. L., de Carvalho, A. C., Fregnani, J. H., Longatto-Filho, A., Reis, R. M., Scapulatempo-Neto, C., Silveira, H. C., Vidal, D. O., Burnette, A., Eschbacher, J., Hermes, B., Noss, A., Singh, R., Anderson, M. L., Castro, P. D., Ittmann, M., Huntsman, D., Kohl, B., Le, X., Thorp, R., Andry, C., Duffy, E. R., Lyadov, V., Paklina, O., Setdikova, G., Shabunin, A., Tavobilov, M., McPherson, C., Warnick, R., Berkowitz, R., Cramer, D., Feltmate, C., Horowitz, N., Kibel, A., Muto, M., Raut, C. P., Malykh, A., Barnholtz-Sloan, J. S., Barrett, W., Devine, K., Fulop, J., Ostrom, Q. T., Shimmel, K., Wolinsky, Y., Sloan, A. E., De Rose, A., Giuliante, F., Goodman, M., Karlan, B. Y., Hagedorn, C. H., Eckman, J., Harr, J., Myers, J., Tucker, K., Zach, L. A., Deyarmin, B., Hu, H., Kvecher, L., Larson, C., Mural, R. J., Somiari, S., Vicha, A., Zelinka, T., Bennett, J., Iacocca, M., Rabeno, B., Swanson, P., Latour, M., Lacombe, L., Têtu, B., Bergeron, A., McGraw, M., Staugaitis, S. M., Chabot, J., Hibshoosh, H., Sepulveda, A., Su, T., Wang, T., Potapova, O., Voronina, O., Desjardins, L., Mariani, O., Roman-Roman, S., Sastre, X., Stern, M. H., Cheng, F., Signoretti, S., Berchuck, A., Bigner, D., Lipp, E., Marks, J., McCall, S., McLendon, R., Secord, A., Sharp, A., Behera, M., Brat, D. J., Chen, A., Delman, K., Force, S., Khuri, F., Magliocca, K., Maithel, S., Olson, J. J., Owonikoko, T., Pickens, A., Ramalingam, S., Shin, D. M., Sica, G., Van Meir, E. G., Zhang, H., Eijckenboom, W., Gillis, A., Korpershoek, E., Looijenga, L., Oosterhuis, W., Stoop, H., van Kessel, K. E., Zwarthoff, E. C., Calatozzolo, C., Cuppini,

L., Cuzzubbo, S., DiMeco, F., Finocchiaro, G., Mattei, L., Perin, A., Pollo, B., Chen, C., Houck, J., Lohavanichbutr, P., Hartmann, A., Stoehr, C., Stoehr, R., Taubert, H., Wach, S., Wullich, B., Kycler, W., Murawa, D., Wiznerowicz, M., Chung, K., Edenfield, W. J., Martin, J., Baudin, E., Bubley, G., Bueno, R., De Rienzo, A., Richards, W. G., Kalkanis, S., Mikkelsen, T., Noushmehr, H., Scarpace, L., Girard, N., Aymerich, M., Campo, E., Giné, E., Guillermo, A. L., Van Bang, N., Hanh, P. T., Phu, B. D., Tang, Y., Colman, H., Evason, K., Dottino, P. R., Martignetti, J. A., Gabra, H., Juhl, H., Akeredolu, T., Stepa, S., Hoon, D., Ahn, K., Kang, K. J., Beuschlein, F., Breggia, A., Birrer, M., Bell, D., Borad, M., Bryce, A. H., Castle, E., Chandan, V., Cheville, J., Copland, J. A., Farnell, M., Flotte, T., Giama, N., Ho, T., Kendrick, M., Kocher, J. P., Kopp, K., Moser, C., Nagorney, D., O'Brien, D., O'Neill, B. P., Patel, T., Petersen, G., Que, F., Rivera, M., Roberts, L., Smallridge, R., Smyrk, T., Stanton, M., Thompson, R. H., Torbenson, M., Yang, J. D., Zhang, L., Brimo, F., Ajani, J. A., Angulo Gonzalez, A. M., Behrens, C., Bondaruk, J., Broaddus, R., Czerniak, B., Esmaeli, B., Fujimoto, J., Gershenwald, J., Guo, C., Lazar, A. J., Logothetis, C., Meric-Bernstam, F., Moran, C., Ramondetta, L., Rice, D., Sood, A., Tamboli, P., Thompson, T., Troncoso, P., Tsao, A., Wistuba, I., Carter, C., Haydu, L., Hersey, P., Jakrot, V., Kakavand, H., Kefford, R., Lee, K., Long, G., Mann, G., Quinn, M., Saw, R., Scolyer, R., Shannon, K., Spillane, A., Stretch, J., Synott, M., Thompson, J., Wilmott, J., Al-Ahmadie, H., Chan, T. A., Ghossein, R., Gopalan, A., Levine, D. A., Reuter, V., Singer, S., Singh, B., Tien, N. V., Broudy, T., Mirsaidi, C., Nair, P., Drwiega, P., Miller, J., Smith, J., Zaren, H., Park, J. W., Hung, N. P., Kebebew, E., Linehan, W. M., Metwalli, A. R., Pacak, K., Pinto, P. A., Schiffman, M., Schmidt, L. S., Vocke, C. D., Wentzensen, N., Worrell, R., Yang, H., Moncrieff, M., Goparaju, C., Melamed, J., Pass, H., Botnariuc, N., Caraman, I., Cernat, M., Chemencedji, I., Clipca, A., Doruc, S., Gorincioi, G., Mura, S., Pirtac, M., Stancul, I., Tcaciuc, D., Albert, M., Alexopoulou, I., Arnaout, A., Bartlett, J., Engel, J., Gilbert, S., Parfitt, J., Sekhon, H., Thomas, G., Rassl, D. M., Rintoul, R. C., Bifulco, C., Tamakawa, R., Urba, W., Hayward, N., Timmers, H., Antenucci, A., Facciolo, F., Grazi, G., Marino, M., Merola, R., de Krijger, R., Gimenez-Roqueplo, A. P., Piché, A., Chevalier, S., McKercher, G., Birsoy, K., Barnett, G., Brewer, C., Farver, C., Naska, T., Pennell, N. A., Raymond, D., Schilero, C., Smolenski, K., Williams, F., Morrison, C., Borgia, J. A., Liptay, M. J., Pool, M., Seder, C. W., Junker, K., Omberg, L., Dinkin, M., Manikhas, G., Alvaro, D., Bragazzi, M. C., Cardinale, V., Carpino, G., Gaudio, E., Chesla, D., Cottingham, S., Dubina, M., Moiseenko, F., Dhanasekaran, R., Becker, K. F., Janssen, K. P., Slotta-Huspenina, J., Abdel-Rahman, M. H., Aziz, D., Bell, S., Cebulla, C. M., Davis, A., Duell, R., Elder, J. B., Hilty, J., Kumar, B., Lang, J., Lehman, N. L., Mandt, R., Nguyen, P., Pilarski, R., Rai, K., Schoenfield, L., Senecal, K., Wakely,

P., Hansen, P., Lechan, R., Powers, J., Tischler, A., Grizzle, W. E., Sexton, K. C., Kastl, A., Henderson, J., Porten, S., Waldmann, J., Fassnacht, M., Asa, S. L., Schadendorf, D., Couce, M., Graefen, M., Huland, H., Sauter, G., Schlomm, T., Simon, R., Tennstedt, P., Olabode, O., Nelson, M., Bathe, O., Carroll, P. R., Chan, J. M., Disaia, P., Glenn, P., Kelley, R. K., Landen, C. N., Phillips, J., Prados, M., Simko, J., Smith-McCune, K., VandenBerg, S., Roggin, K., Fehrenbach, A., Kendler, A., Sifri, S., Steele, R., Jimeno, A., Carey, F., Forgie, I., Mannelli, M., Carney, M., Hernandez, B., Campos, B., Herold-Mende, C., Jungk, C., Unterberg, A., von Deimling, A., Bossler, A., Galbraith, J., Jacobus, L., Knudson, M., Knutson, T., Ma, D., Milhem, M., Sigmund, R., Godwin, A. K., Madan, R., Rosenthal, H. G., Adebamowo, C., Adebamowo, S. N., Boussioutas, A., Beer, D., Giordano, T., Mes-Masson, A. M., Saad, F., Bocklage, T., Landrum, L., Mannel, R., Moore, K., Moxley, K., Postier, R., Walker, J., Zuna, R., Feldman, M., Valdivieso, F., Dhir, R., Luketich, J., Mora Pinero, E. M., Quintero-Aguilo, M., Carlotti, C. G., Dos Santos, J. S., Kemp, R., Sankarankuty, A., Tirapelli, D., Catto, J., Agnew, K., Swisher, E., Creaney, J., Robinson, B., Shelley, C. S., Godwin, E. M., Kendall, S., Shipman, C., Bradford, C., Carey, T., Haddad, A., Moyer, J., Peterson, L., Prince, M., Rozek, L., Wolf, G., Bowman, R., Fong, K. M., Yang, I., Korst, R., Rathmell, W. K., Fantacone-Campbell, J. L., Hooke, J. A., Kovatich, A. J., Shriver, C. D., DiPersio, J., Drake, B., Govindan, R., Heath, S., Ley, T., Van Tine, B., Westervelt, P., Rubin, M. A., Lee, J. I., Aredes, N. D., Mariamidze, A., Buonamici, S., et Yu, L. (2018) Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Reports*, 23(1) : 282–296.e4.

Simoneau, J., Dumontier, S., Gosselin, R., et Scott, M. S. (2019) Current RNA-seq methodology reporting limits reproducibility. *Briefings in Bioinformatics*.

Simoneau, J., Gosselin, R., et Scott, M. S. (2020) Factorial study of the RNA-seq computational workflow identifies biases as technical gene signatures. *bioRxiv*.

Spengler, S. J. (2000) Bioinformatics in the information age. `https://science.sciencemag.org/content/287/5456/1221`.

Stein, L. (2001) Genome annotation : From sequence to biology. *Nature Reviews Genetics*, 2(7) : 493–503.

Stein, L. (2013) Generic Feature Format Version 3 (GFF3). `https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md`.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott,

P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., et Collins, R. (2015) UK Biobank : An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3) : 1–10.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., et Yaspo, M. L. (2014) Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics*, 15(1).

Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., Li, S., Mason, C. E., Olson, S., Pervouchine, D., Sloan, C. A., Wei, X., Zhan, L., et Irizarry, R. A. (2016) A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17(1).

Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M., et Kitts, P. (2013) Eukaryotic Genome Annotation Pipeline. *The NCBI Handbook [Internet]. 2nd edition.*

Thierry-Mieg, D. et Thierry-Mieg, J. (2006) AceView : a comprehensive cDNA-supported gene and transcripts annotation. *Genome biology*, 7 Suppl 1(Suppl 1) : 1–14.

Trombetta, J., Gennert, D., Lu, D., et Sattija, R. (2015) Preparation of single-cell RNA-seq libraries for NGS. *Curr Protoc Mol Biol*, 19(2) : 161–169.

UCSC Genome Browser (2015) New default gene set on GRCh38 : GENCODE Basic genes. `http://genome.ucsc.edu/blog/new-default-gene-set-on-grch38-gencode-basic-genes/`.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metzstein, M., Hawkins, T., Wilson, R., Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J., et Sulston, J. (1992) A survey of expressed genes in Caenorhabditis elegans. *Nature Genetics*, 1(may).

Wellcome Sanger Institute (2019) HAVANA Manual Annotation. `https://www.sanger.ac.uk/science/projects/manual-annotation`.

Wery, M., Descrimes, M., Thermes, C., Gautheret, D., et Morillon, A. (2013) Zinc-mediated RNA fragmentation allows robust transcript reassembly upon whole transcriptome RNA-Seq. *Methods*, 63(1) : 25–31.

Williams, C. R., Baccarella, A., Parrish, J. Z., et Kim, C. C. (2016) Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*, 17(1) : 1–13.

Williams, C. R., Baccarella, A., Parrish, J. Z., et Kim, C. C. (2017) Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, 18(1) : 1–13.

Wilming, L. G., Gilbert, J. G., Howe, K., Trevanion, S., Hubbard, T., et Harrow, J. L. (2008) The vertebrate genome annotation (VEGA) database. *Nucleic Acids Research*, 36(SUPPL. 1) : 753–760.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Soulette, C. M., Snutch, T. P., Loman, N., Paten, B., Loose, M., Simpson, J. T., Olsen, H. E., Brooks, A. N., Akeson, M., et Timp, W. (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature methods*, 16(December).

Wu, P.-Y., Phan, J. H., et Wang, M. D. (2013) Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics*, 14(Suppl 11) : S8.

Yamasaki, C., Murakami, K., ichi Takeda, J., Sato, Y., Noda, A., Sakate, R., Habara, T., Nakaoka, H., Todokoro, F., Matsuya, A., Imanishi, T., et Gojobori, T. (2009) H-InvDB in 2009 : Extended database and data mining resources for human genes and transcripts. *Nucleic Acids Research*, 38(SUPPL.1) : 626–632.

Yates, B., Braschi, B., Gray, K. A., Seal, R. L., Tweedie, S., et Bruford, E. A. (2017) Genenames.org : The HGNC and VGNC resources in 2017. *Nucleic Acids Research*, 45(D1) : D619–D625.

Zhang, Y., Gong, M., Yuan, H., Park, H. G., Frierson, H. F., et Li, H. (2012) Chimeric transcript generated by cis- splicing of adjacent genes regulates prostate cancer cell proliferation. *Cancer Discovery*, 2(7) : 598–607.

Zhang, Y., Yang, L., Kucherlapati, M., Chen, F., Hadjipanayis, A., Pantazi, A., Bristow, C. A., Lee, E. A., Mahadeshwar, H. S., Tang, J., Zhang, J., Seth, S., Lee, S., Ren, X., Song, X., Sun, H., Seidman, J., Luquette, L. J., Xi, R., Chin, L., Protopopov, A., Li, W., Park, P. J., Kucherlapati, R., et Creighton, C. J. (2018) A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Reports*, 24(2) : 515–527.